

Perfect Bayesian Equilibria in Sequential Reputation Games *

Nuh Aygün Dalkıran[†]

Serdar Yüksel^{‡§}

December 5, 2016

Abstract

We analyze reputation games where a strategic long-lived player acts in a sequential repeated game against a collection of short-lived players. The key assumption in our model is that the information of the short-lived players is nested in that of the long-lived player. Under this assumption,

(i) We show that, given mild assumptions, the set of Perfect Bayesian Equilibrium payoffs coincide with Markov Perfect Equilibrium payoffs in the standard discounted average payoff setup.

(ii) A dynamic programming formulation can be obtained for the computation of equilibrium strategies of the strategic long-lived player in the standard discounted average payoff setup.

(iii) We consider the undiscounted average payoff setup separately where we obtain an optimal equilibrium strategy of the strategic long-lived player.

(iv) We then use this optimal strategy in the undiscounted average payoff setup as a tool to obtain an upper payoff bound for the arbitrarily patient long-lived player in the standard discounted average payoff setup.

(v) By using measure concentration techniques, we obtain a refined lower payoff bound on the value of reputation in the standard discounted average payoff setup.

(vi) Under further assumptions, we establish the continuity of the equilibrium payoffs, in the standard discounted average payoff setup.

Keywords: Stochastic Control, Reputations, Repeated Games, Incomplete Information, Long and Short-lived Players.

JEL classification: C61, C70, C72, C73, D83

*We are grateful to Tamás Linder for his collaborations on some of the results that led to part of the contributions here. We would like to thank Daron Acemoglu, Alp Atakan, Mehmet Ekmekci, Olivier Gossner, Johannes Hörner and Bruno Jullien for helpful discussions and the seminar participants at Koç University, Toulouse School of Economics, and the participants of the 2nd Occasional Workshop in Economic Theory at University of Graz, the 69th European Meeting of the Econometric Society, Geneva, Switzerland, and the 11th World Congress of the Econometric Society, Montreal, Canada for helpful suggestions. We also thank Oral Ersoy Dokumacı for research assistance.

[†]Department of Economics, Bilkent University, Çankaya, Ankara, 06800, Turkey; email: dalkiran@bilkent.edu.tr

[‡]Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, Canada, K7L 3N6; email: yuk-sel@mast.queensu.ca

[§]This research was partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Scientific and Technological Research Council of Turkey (TUBITAK).

1 Introduction

It is well known that “reputation” plays an important role in long run relationships. When one considers buying a product from a particular firm, his action will depend on his belief about this firm, i.e., the firm’s reputation, which he has formed based on previous experiences. Many interactions among economic agents are repeated and are in the form of long-run relationships. Therefore, economists have been extensively studying the role of reputation in long-run relationships and repeated games.

By using reputations as a conceptual as well as a mathematical quantitative variable, economists have been able to explain how reputation can rationalize the intuitive equilibria, as in the expectation of cooperation in early rounds of a finitely repeated prisoners’ dilemma (Kreps et al. [39]), and entry deterrence in the early rounds of the chain store game (Kreps and Wilson [40], Milgrom and Roberts [45]). Reputational concerns can also explain the benefit of providing high quality products (Klein and Leffler [38]), and the benefit of generating good returns to investors (Diamond [17]). Reputation can help time-inconsistent governments to commit to non-inflationary monetary policies (Barro[5], Cukierman and Meltzer[15]), low capital taxation (Chari and Kehoe [11], Celentani and Pesendorfer[10]), and repayment of sovereign debt (Cole, Dow, and English [12]).

Recently, there has been an emergence of use of tools from information and control theory in the reputations literature (see Gossner [30], Ekmekci, Gossner, and Wilson [20], Faingold[22]). Such tools have been proved to be useful in studying the bounds on the value of reputation.

In this paper, by adopting and generalizing recent results from stochastic control theory, we provide a new approach and establish refined results on repeated games with incomplete information. Before stating our contributions and the problem setup more explicitly, we provide a brief overview of the related literature in the following subsection.

1.1 Related Literature

Kreps, Milgrom, Roberts, and Wilson (see [39], [40] and [45]) introduced the adverse selection approach to study reputations in (finitely) repeated games. Fudenberg and Levine [25], [26] extended this approach to infinitely repeated games and showed that a patient long-lived player facing infinitely many short-lived players can guarantee himself a payoff close to his Stackelberg payoff when there is a slight probability that the long-lived player is a commitment type who always plays the stage game Stackelberg action. When compared to the folk theorem (see Fudenberg and Maskin[29], Fudenberg, Levine, and Maskin[28]), their results imply an intuitive expectation: the equilibria with relatively high payoffs are more likely to arise due to reputation effects.

Even though the results of Fudenberg and Levine [25], [26] hold for both perfect and imperfect public monitoring, Cripps, Mailath and Samuelson [14] showed that reputation effects are not sustainable in the long-run when there is imperfect public monitoring. In other words, under imperfect public monitoring it is impossible to maintain a permanent reputation for playing a strategy that does not play an equilibrium of the complete information game.

Since Cripps, Mailath, and Samuelson’s work [14], there has been a large literature which studies the possibility / impossibility of maintaining permanent reputations: Ekmekci [19]

showed that reputation can be sustained permanently in the steady state by using rating systems. Ekmekci, Gossner, and Wilson [20] showed that impermanent types could lead to permanent reputations as well. Atakan and Ekmekci [2], [3], [4] and Özdoğan [46] provided positive and negative results on permanent reputations with long-lived players on both sides. Liu [42] provided dynamics that explain accumulation, consumption, and restoration of reputation when the discovery of the past is costly. Liu and Skrzypacz [43] provided similar dynamics for when there is limited record-keeping. Faingold and Sannikov [23] studied continuous time reputation games and provided a characterization of the equilibrium payoff set by making use of stochastic differential equations. Hörner and Lovo [34] analyzed reputations by making use of the concept of belief free equilibria in repeated games with incomplete information.

Sorin [51] unified and improved some of the results in reputations literature by using tools from Bayesian learning and merging due to Kalai and Lehrer [36], [37]. Gossner [30] was the first to point out that making use of the concept of relative entropy has advantages in terms of tractability to study bounds on the value of reputation: He provided powerful bounds on the value of reputations by employing basic properties (such as the chain rule) of relative entropy. These bounds coincide in the limit (as the strategic long-lived player becomes arbitrarily patient) with the bounds provided by Fudenberg and Levine [25], [26]. Faingold [22] made use of these techniques to provide reputation bounds for repeated moral hazard games in which a long-lived player interacts frequently with a population of short-lived players where the monitoring technology varies with the length of the period of interaction.

In economics, entropy and information theoretic techniques were not only used in the reputations literature; Sims [49], [50] used the concept of mutual information which uses the reduction in the entropy of a random variable as the measure of information flow. This started a new avenue in Macroeconomics on rational inattention.

1.2 Contributions and Connections with the Literature

Contributions of the paper. Our findings contribute to the reputations literature by obtaining structural and computational results on the equilibrium behavior in finite-horizon, infinite-horizon, and undiscounted settings in sequential reputation games, as well as refined upper and lower bounds on the value of reputations: We analyze reputation games where a strategic long-lived player acts in a repeated sequential-move game against a collection of short-lived players each of whom plays the stage game only once but observes signals correlated with interactions of the previous short-lived players. The key assumption in our model is that the information of the short-lived players is nested in that of the long-lived player in a causal fashion. That is, any information available to a short-lived player in some period becomes available to the long-lived player in the following period. This nested information structure is obtained through an appropriate almost standard monitoring structure. Under this monitoring structure, we obtain stronger results than what currently exists in the literature in a number of directions.

In particular:

- (i) Given mild assumptions, we show that the set of Perfect Bayesian Equilibrium payoffs coincide with Markov Perfect Equilibrium payoffs.
- (ii) Hence, a dynamic programming formulation is obtained for the computation of

equilibrium strategies of the strategic long-lived player in the standard discounted average payoff setup.

(iii) In the undiscounted average payoff setup, under an identifiability assumption, we obtain an optimal -equilibrium- strategy for the strategic long-lived player. To the best of our knowledge, the undiscounted setup has not been studied in the reputations literature so far. In particular, we provide new techniques to investigate the optimality of mimicking a Stackelberg commitment type in the undiscounted average payoff setup.

(iv) The optimal strategy we obtain in the undiscounted average payoff setup also lets us obtain, through an Abelian inequality, an upper payoff bound for the arbitrarily patient long-lived player –in the discounted average payoff setup. We show that this achievable upper bound is identified with a stage game Stackelberg equilibrium payoff.

(v) By using measure concentration techniques, we obtain a refined lower payoff bound on the value of reputation for a fixed discount factor. This lower bound coincide with the lower bounds identified by Fudenberg and Levine [26] and Gossner [30] as the long-lived player becomes arbitrarily patient, i.e., as his discount factor tends to 1.

(vi) Finally, under further assumptions, we establish the continuity of the equilibrium payoffs in the standard discounted average payoff setup.

Connections with related literature. There are a number of related results in the information theory and control literature on real-time signaling which provide powerful structural, topological, and operational results that are in principle similar to the reputations models analyzed in economics, despite the simplifications that come about due to the fact that in these fields, the players typically have a common utility function. Furthermore, such studies typically assume finitely repeated setups, whereas we consider here infinitely repeated setups which require non-trivial generalizations. We particularly mention Yüksel [56], Linder and Yüksel [41], Borkar, Mitter, and Tatikonda [9], Yüksel and Başar [57], Witsenhausen[55], Walrand and Varaiya [54], Teneketzis [53] and Mahajan and Teneketzis [44] for various contexts but note that all of these studies except Borkar, Mitter, and Tatikonda [9] and Linder and Yüksel [41] have focused on finite horizon problems. Using tools of stochastic control theory and zero-delay source coding, we provide new techniques to study reputations: These techniques not only result in a number of conclusions re-affirming certain results documented in the reputations literature, but also provide new results and interpretations as we briefly discuss in the following.

Our first set of results are regarding the optimal strategies of the long-lived player: We show that when the information available to the short-lived players is also available to the long-lived player, i.e., when the information structure is nested, given an arbitrary fixed sequence of strategies of the short-lived players, for any (private) strategy of the long-lived player, there exists a public strategy (a strategy that depends only on short-lived players' information) which performs at least as good as the original (private) strategy against this fixed sequence of strategies of the short-lived players. We show that this is true for both finite horizon and infinite horizon games and against both for (Bayesian) rational and non-rational strategies of the short-lived players. Hence, these results might be useful for the study of reputation against non-Bayesian and/or boundedly rational (short-lived) players.

On the other hand, when the short-lived players respond optimally in a (Bayesian) rational fashion, their posterior beliefs serve as a sufficient statistic of their information. Building on this we show, under a mild measurability assumption on the short-lived players'

strategies, that Perfect Bayesian equilibrium payoffs and Perfect Public equilibrium payoffs coincide with each other. A relevant result is due to Fudenberg and Levine [27] who show that sequential equilibrium payoffs and perfect public equilibrium payoffs coincide in a similar but different setup. Not only our setup is different than that of Fudenberg and Levine [27] but also our proof techniques use results from stochastic control theory which leads to more general results that hold for finite horizon and allow for non-equilibrium strategies for short-lived players as well.

Building on these results, our second main result is to model the optimization problem the strategic long-lived player faces as a controlled Markov chain to obtain his equilibrium behavior. We show that the strategic long-lived player's problem can be solved through an infinite horizon discounted payoff dynamic programming method. These results imply a stationary (time-invariant given the posteriors) Markov Perfect Equilibrium. Hence, we obtain the equivalence of Perfect Bayesian Equilibrium payoffs and Markov Perfect Equilibrium payoffs of the strategic long-lived player. Furthermore, we establish the continuity of any equilibrium payoff value in the prior of the short-lived players for every fixed discount factor, under a technical condition. This result is in agreement with the findings of Dalkiran [16] who uses methods of Abreu, Pearce, and Stacchetti [1] to show similar continuity results and implies that a conjecture by Cripps, Mailath, and Samuelson [14] is indeed true under this technical condition. We note that the controlled Markovian construction is also aligned with the observation that the set of Perfect Bayesian Equilibria of an infinitely repeated incomplete information game coincide with a family of Markov chains which was made by Bergin [6].

Our third set of results are about the undiscounted average payoff setup which help us provide an upper payoff bound for the value of reputation for an arbitrarily patient long-lived player in the standard discounted average payoff setup: This setup has not been studied in the reputations literature. Here, we show, under an identifiability assumption, that an optimal stationary equilibrium strategy for the strategic-long-lived player in the undiscounted average payoff setup is to mimic a Stackelberg commitment type forever. This optimal equilibrium strategy combined with an Abelian inequality delivers that *an upper payoff bound for the value of reputation for an arbitrarily patient long-lived player in the discounted average payoff setup* is a stage game Stackelberg payoff.

Our fourth set of results are about providing a lower payoff bound for the value of reputation: Through an explicit measure-concentration analysis, we identify a lower payoff bound for the value of reputation for a fixed discount factor. This lower payoff bound is stronger than the lower payoff bounds identified by Fudenberg and Levine [26] and Gossner [30] *for a fixed discount factor*. Yet, in the limit as the discount factor tends to 1, our lower payoff bound converges to a stage game Stackelberg payoff as in the case of Fudenberg and Levine [26] and Gossner [30] – when the long-lived player has a commitment type which always plays the associated stage game Stackelberg action. That is, in agreement with the previous findings, we observe that if the long-lived player has a commitment type which plays a stage game Stackelberg action, then no matter how small the prior measure on this type is, a payoff which is arbitrarily close to the associated Stackelberg payoff can be achieved by a sufficiently patient long-lived player. Thus, we obtain parallel results to those of Fudenberg and Levine [26] and Gossner [30] through obtaining lower bounds and upper

bounds for the value of reputation by employing novel techniques.¹ We believe that these novel techniques will be useful to those who work on similar problems in the reputations literature.

In the next section, we present preliminaries of our model as well as a motivating example. Section 3 provides our structural results leading to the equivalence of Perfect Bayesian Equilibrium payoffs and Markov Perfect Equilibrium payoffs in the standard discounted average payoff setup. Section 4 provides results characterizing the optimal behavior of the long-lived player for the undiscounted average payoff setup which lead us to an upper bound for the equilibrium payoffs in the standard discounted average payoff setup when the long-lived player becomes arbitrarily patient. Section 5 provides, through an explicit measure concentration analysis, a lower bound for the equilibrium payoffs of the strategic long-lived player under reputation concerns in the standard discounted average payoff setup. Section 6 provides the continuity of the equilibrium payoffs in the standard discounted average payoff setup. Section 7 concludes the paper. We also provide a brief review of Markov Decision Processes at the end of the Appendix.

2 The Model

A long-lived player (Player 1) plays a repeated stage game with a sequence of different short-lived players (Player 2). Action sets of Player 1 and Player 2 in the stage game are assumed to be finite and denoted by \mathbb{A}^1 and \mathbb{A}^2 , respectively.

There is incomplete information regarding the type of the long-lived Player 1. The set of all possible types of Player 1 is given by $\Omega = \{\omega^n\} \cup \hat{\Omega}$ where ω^n is the normal (strategic) type and $\hat{\Omega}$ is a finite set of commitment types. Each type $\hat{\omega} \in \hat{\Omega}$ is a simple type committed to playing the corresponding (possibly mixed) action $\hat{\omega} \in \Delta(\mathbb{A}^1)$ at every stage of the interaction independent of the history of the play.² The *common knowledge* initial prior over Player 1's types, $\mu_0 \in \Delta(\Omega)$, is assumed to have full support.

The stage game is a *sequential-move* game: Player 1 moves first; when action a^1 is chosen by Player 1 in the stage game; a **public** signal $z^2 \in \mathbb{Z}^2$ is observed by Player 2 which is drawn according to the probability distribution $\rho^2(\cdot|a^1) \in \Delta(\mathbb{Z}^2)$. Player 2, observing his signal, moves second. At the end of the stage game, Player 1 observes a **private** signal $z^1 \in \mathbb{Z}^1$ which depends on actions of both players and is drawn according to the probability distribution $\rho^1(\cdot|(a^1, a^2))$. Both the set of Player 1's all possible private signals, \mathbb{Z}^1 , and the set of Player 2's all possible public signals, \mathbb{Z}^2 , are assumed to be finite.

There is a **nested information structure** in this repeated game in the following sense:

¹We note that our results regarding lower and upper payoff bounds focus on the concept of Perfect Bayesian Equilibrium, whereas the aforementioned results in the literature due to Fudenberg and Levine [26] and Gossner [30] are also applicable to the Bayes-Nash equilibrium concept. These two concepts coincide with each other when the monitoring structure satisfies a full support assumption as in Cripps, Mailath and Samuelson [14], i.e., when each stage game signal realizes with positive probability under any stage game action profile. In such a case, any finite sequence of signals realizes with positive probability therefore must be followed by optimal behavior in any Nash Equilibrium. The only out of equilibrium paths will be those in which a short-lived player deviates but in a Nash equilibrium such a deviation will not be profitable. Hence, any Perfect Bayesian Equilibrium outcome will also be a Nash equilibrium outcome. We note that most of our results will not require such a full support monitoring assumption.

² $\Delta(\mathbb{A}^1)$ denotes the set of all probability measures on \mathbb{A}^1 .

The signals observed by Player 2s are **public** and hence available to all subsequent players whereas Player 1's signals are his **private** information. Therefore, **the information of Player 2 at time $t - 1$ is a subset of the information of Player 1 at time t** . In the rest of the paper, we refer this monitoring structure as a **nested information structure** to imply that **the information of Player 2s is nested in that of Player 1**.

Formally, a generic history for Player 2 at time $t - 1$ and a generic history for Player 1 at time t are given as follows:

$$h_{t-1}^2 = (z_0^2, z_1^2, \dots, z_{t-1}^2) \in (\mathbb{Z}^2)^t =: H_{t-1}^2 \quad (1)$$

$$h_t^1 = (a_0^1, z_0^1, z_0^2, \dots, a_{t-1}^1, z_{t-1}^1, z_{t-1}^2) \in (\mathbb{A}^1 \times \mathbb{Z}^1 \times \mathbb{Z}^2)^t =: H_t^1 \quad (2)$$

That is, each Player 2 observes, before he acts, a finite sequence of public signals which are correlated with Player 1's interaction with previous Player 2s. On the other hand, Player 1 observes not only these public signals, but also a sequence of private signals for each particular interaction that happened in the past, and his actions in the previous periods.³

Due to its importance for our results, we explicitly note this nested information structure as Remark 2.1.

Remark 2.1 (Nested Information). The signals observed by Player 2s are public and hence available to all subsequent players whereas Player 1's signals are his private information.

We note also that having such a monitoring structure is not a strong assumption. In particular, it is weaker than the information structure in Fudenberg and Levine [26] where it is assumed that only the same sequence of public signals are observable by the long-lived and short-lived players, i.e., there is only public monitoring. Yet, it is stronger than the information structure in Gossner [30] which allows private monitoring for both the long-lived and short lived players. Therefore, the nested information structure we consider falls somewhere in between Fudenberg and Levine [26] and Gossner [30] in terms of monitoring structures.⁴

The stage game payoff function of the normal (strategic) type long-lived Player 1 is given by u^1 , and each short-lived Player 2's payoff function is given by u^2 , where $u^i : \mathbb{A}^1 \times \mathbb{A}^2 \rightarrow \mathbb{R}$.

The set of all possible histories for Player 2 of stage t is $H_t^2 = H_{t-1}^2 \times \mathbb{Z}^2$ where $H_{t-1}^2 = (\mathbb{Z}^2)^t$ as was defined in (1). On the other hand, the set of all possible histories observable by the long-lived Player 1 prior to stage t is $H_t^1 = (\mathbb{A}^1 \times \mathbb{Z}^1 \times \mathbb{Z}^2)^t$ as was defined in (2). It is assumed that $H_0^1 := \emptyset$ and $H_0^2 := \emptyset$, which is the usual convention. Let $\mathcal{H}^1 = \bigcup_{t \geq 0} H_t^1$ be the set of all possible histories of the long-lived Player 1.

A (behavioral) strategy for Player 1 is a map:

$$\sigma^1 : \Omega \times \mathcal{H}^1 \rightarrow \Delta(\mathbb{A}^1).$$

that satisfies $\sigma^1(\hat{\omega}, h_{t-1}^1) = \hat{\omega}$ for any $\hat{\omega} \in \hat{\Omega}$ and for every $h_{t-1}^1 \in H_{t-1}^1$, since commitment types are required to play the same corresponding strategy of the stage game independent of the history. The set of all strategies for Player 1 is denoted Σ^1 .

³Note that Player 1 gets to observe the realizations of his earlier mixed actions.

⁴Our model allows for perfect monitoring and imperfect public monitoring as special cases as well.

A strategy for Player 2 of stage t is a map:

$$\sigma_t^2 : H_{t-1}^2 \times \mathbb{Z}^2 \rightarrow \Delta(\mathbb{A}^2).$$

We let Σ_t^2 be the set of all such strategies and let $\Sigma^2 = \prod_{t \geq 0} \Sigma_t^2$ denote the set of all sequences of all such strategies. A history (or path) h_t of length t is an element of $\Omega \times (\mathbb{A}^1 \times \mathbb{A}^2 \times \mathbb{Z}^1 \times \mathbb{Z}^2)^t$ describing Player 1's type, actions, and signals realized up to stage t . By standard arguments in the field (e.g. Kolmogorov's Extension Theorem (see Hernandez-Lerma and Lasserre [31])), a strategy profile $\sigma = (\sigma^1, \sigma^2) \in \Sigma^1 \times \Sigma^2$ induces a unique probability distribution P_σ over the set of all paths of play $H^\infty = \Omega \times (\mathbb{A}^1 \times \mathbb{A}^2 \times \mathbb{Z}^1 \times \mathbb{Z}^2)^\mathbb{N}$ endowed with the product σ -algebra.

We let $a_t = (a_t^1, a_t^2)$ represent the action profile realized at stage t and let $z_t = (z_t^1, z_t^2)$ denote the signal profile realized at stage t . Given $\omega \in \Omega$, $P_{\omega, \sigma}(\cdot) = P_\sigma(\cdot | \omega)$ represents the probability distribution over all paths of play conditional on Player 1 being type ω . Player 1's discount factor is assumed to be $\delta \in (0, 1)$ and hence, the expected discounted average payoff to the strategic (normal type) long-lived Player 1 is given by

$$\pi_1(\sigma) = \mathbb{E}_{P_{\omega^n, \sigma}}(1 - \delta) \sum_{t \geq 0} \delta^t u^1(a_t).$$

In most of our results, we will assume that Player 2s are Bayesian rational.⁵ Hence, we will restrict attention to Perfect Bayesian Nash equilibrium: In any such equilibrium, the strategic Player 1 maximizes his expected discounted average payoff given that the short-lived players play a best response to their expectations according to their updated beliefs.⁶ Each Player 2, playing the stage game only once, will be best-responding to his expectation according to his beliefs which are updated according to the Bayes' Rule.

A strategy of Player 2s, σ^2 , is a best response to σ^1 if, for all t ,

$$\mathbb{E}_{P_\sigma}[u^2(a_t^1, a_t^2) | z_{[0, t]}^2] \geq \mathbb{E}_{P_\sigma}[u^2(a_t^1, a^2) | z_{[0, t]}^2] \text{ for all } a^2 \in A^2 \text{ } P_\sigma - a.s.$$

where $z_{[0, t]}^2 = (z_0^2, z_1^2, \dots, z_t^2)$ denotes the information available to Player 2 at time t .

2.1 Motivating Example: A Consultant with Reputation Concerns under Moral Hazard

We provide below a simple example which motivates our model:

A free-lance consultant is to advise different firms in different projects. In each of these projects, a supervisor from the particular firm is to inspect the consultant regarding his effort during the particular project. The consultant can either exert a (H)igh level of effort or a (L)ow level of effort while working on the project.

There is a moral hazard problem: The effort of the management consultant is not directly observable to the supervisor. Yet, after the management consultant chooses his effort level,

⁵A Bayesian rational Player 2 tries to maximize his expected payoff after updating his beliefs according to the Bayes' rule whenever possible. Some of our structural results on equilibrium behavior does not require Bayesian rationality and holds for non-Bayesian Player 2s who might underreact or overreact to new (or recent) information as in Epstein et al. [18] as well.

⁶This will be appropriately modified when we consider the undiscounted average payoff setup.

the supervisor gets to observe a public signal $z^2 \in \{h, l\}$ which is correlated with the effort level of the consultant according to the probability distribution $\rho^2(h|H) = \rho^2(l|L) = p > \frac{1}{2}$.

Observing this public signal, the supervisor recommends to the upper administration to give the consultant a (B)onus or (N)ot. The consultant does not observe the supervisor's recommendation but a private signal $z^1 \in \{b, n\}$ which is correlated both with his effort level and the supervisor's recommendation according to the following probability distribution: $\rho^1(b|(H, B)) = \rho^1(n|(L, N)) = q > \rho^1(b|(H, N)) = \rho^1(n|(L, B)) = r > \frac{1}{2}$. That is, exerting (H)igh level of effort increases the chance of getting the bonus –for a fixed action of the supervisor, i.e., $\rho^1(b|(H, \cdot)) > \rho^1(b|(L, \cdot))$. Similarly, the probability of getting the bonus is higher when the supervisor recommends the (B)onus –for a fixed action of the consultant, i.e., $\rho^1(b|(\cdot, B)) > \rho^1(b|(\cdot, N))$.

The supervisor prefers to recommend a (B)onus when the consultant works hard (exerts (H)igh effort) and (N)ot to recommend a bonus when the consultant shirks (exerts (L)ow effort). For the management consultant exerting a high level of effort is costly:⁷

	B	N
H	1, 1	-1, -1
L	2, -2	0, 0

It is commonly known that there is a positive probability $p_0 > 0$ with which the consultant is an honorable consultant who always exerts (H)igh level of effort. That is, with $p_0 > 0$ probability the consultant is a *commitment type* who plays H at every period of the repeated game independent of the history.

Consider the incentives of a strategic (non-commitment or normal type) consultant: Does such a consultant have an incentive to build a reputation by exerting a high level effort, if the game is repeated only finitely many times? What kind of equilibrium behavior would one expect from such a consultant if the game is repeated infinitely many times with discounting for a *fixed discount factor*? For example, if he is building a reputation, how often does he shirk (exert (L)ow level of effort)? Does there exist reputation cycles, i.e., does the consultant build a reputation by exerting high effort for a while and then milks it by exerting low effort until his reputation level falls under a particular threshold? What happens when the consultant becomes arbitrarily patient, i.e., his discount factor tends to 1? What can we say about the consultant's optimal reputation building strategy when he does not discount the future but rather cares about his undiscounted average payoff?

The aim of this paper is to provide applied economists with tractable techniques to answer questions similar to the ones mentioned above in settings where economic agents have reputational concerns in a sequential repeated game setup.

3 Optimal Strategies and Equilibrium Behavior

Our first set of results will be regarding the optimal strategies of the strategic long-lived Player 1.

⁷Note that the stage game is a sequential game, the payoffs are summarized in a payoff matrix just for illustrative purposes.

Briefly, since each Player 2 plays the stage game only once, we show that when the information of Player 2 is nested in that of Player 1, under a plausible measurability assumption, the strategic long-lived Player 1 can, without any loss in payoff performance, formulate his strategy as a controlled Markovian system optimization, and thus through dynamic programming. The discounted nature of the optimization problem then leads to the existence of a stationary solution. This implies that for any Perfect Bayesian Equilibrium, there exists a payoff-equivalent stationary Markov Perfect Equilibrium. Hence, we conclude that the Perfect Bayesian Equilibrium payoff set and Markov Perfect Equilibrium payoff set of the strategic long-lived Player 1 coincide with each other.

Below, we provide three results on optimal strategies following steps parallel to Yüksel and Başar [57] which builds on Witsenhausen [55], Walrand and Varaiya [54], Teneketzis [53], and Yüksel [56]. These structural results on optimal strategies will be the key for our first main result, Theorem 3.1, and for the following Markov chain construction.

3.1 Optimal Strategies: Finite Horizon

We first consider the finitely repeated game setup where the stage game is to be repeated $T \in \mathbb{N}$ times. In such a case, the strategic long-lived Player 1 is to maximize $\pi_1(\sigma)$ given by

$$\pi_1(\sigma) = \mathbb{E}_{P_{\omega^n, \sigma}}(1 - \delta) \sum_{t=0}^{T-1} \delta^t u^1(a_t).$$

Our first result, Lemma 3.1, shows that, given any fixed sequence of strategies of the short-lived Player 2s, any optimal strategy of the strategic long-lived Player 1 can be replaced, without any loss in payoff performance, by another optimal strategy which only depends on the (public) information of Player 2s. More specifically, we show that for any **private** strategy of the long-lived Player 1 against an arbitrary sequence of strategies of Player 2s, there exists a **public** strategy of the long-lived Player 1 against the very same sequence of strategies of Player 2s which gives the strategic long-lived player a weakly better payoff.

To the best of our knowledge, this is a new result in the repeated games literature. What is different here from similar results in the repeated games literature is that this is true even when Player 2s strategies are non-Bayesian.⁸

Before we state Lemma 3.1, we note here that the signal z_t^2 that will be available to short-lived Player 2s after round t only depends on the action of the long-lived Player 1 at round t and that the following holds for all $t \geq 1$.

$$P_\sigma(z_t^2 | a_t^1; a_s^1, a_s^2, s \leq t-1) = P_\sigma(z_t^2 | a_t^1). \quad (3)$$

Observation (3) plays an important role in the proof of Theorem 1.

⁸As mentioned before, a relevant result appears in Fudenberg and Levine [27] who show that sequential equilibrium payoffs and perfect public equilibrium payoffs coincide (see Appendix B of Fudenberg and Levine [27]) in a similar infinitely repeated game setup.

Lemma 3.1. *In the finitely repeated setup, given any sequence of strategies of short-lived Player 2s, for any (private) strategy of the strategic long-lived Player 1, there exists a (public) strategy that only conditions on $\{z_0^2, z_1^2, \dots, z_{t-1}^2\}$ which yields the strategic long-lived Player 1 a weakly better payoff against the given sequence of strategies of Player 2s.*

Proof. See Appendix.

Lemma 3.1 implies that any private information of Player 1 is statistically irrelevant for optimal strategies: for any private strategy of the long-lived Player 1, there exists a public strategy which performs at least as good as the original one against a given sequence of strategies of Player 2s. That is, in the finitely repeated setup, the long-lived Player 1 can depend his strategy only on the public information and his type without any loss in payoff performance. We would like to note here once again that Lemma 3.1 above holds for any sequence of strategies of Players 2, even non-Bayesian ones.

On the other hand, when Player 2s are Bayesian rational, as is the norm in repeated games, we obtain a more refined structural result which we state below as Lemma 3.2. As mentioned before, in a Perfect Bayesian Equilibrium the short-lived Player 2 at time t , playing the game only once, seeks to maximize $\sum_{a^1} P_\sigma(a_t^1 | z_{[0,t]}^2) u^2(a^1, a^2)$. However, it may be that his best response set, i.e. the maximizing action set $\arg \max(\sum_{a^1} P_\sigma(a_t^1 | z_{[0,t]}^2) u^2(a^1, a^2))$, may not be unique.

We add the following measurability assumption, which essentially requires that the strategy of Player 2 is a measurable function of their posterior beliefs. Under this assumption, when short-lived Player 2s are Bayesian rational, their posterior beliefs become a sufficient statistic of their information.

Assumption 3.1. *Each Player 2's strategy is a measurable function of $P_\sigma(a_t^1 | z_{[0,t]}^2)$ and t .*

Note that Assumption 3.1 **does not** require that when faced with the same posterior belief each Player 2 must best reply in the same way. It just requires that their strategies must be a measurable function of the posterior belief for each Player 2.⁹

Lemma 3.2. *In the finitely repeated setup, under Assumption 3.1, given any arbitrary sequence of strategies of Bayesian rational short-lived Player 2s, for any (private) strategy of the strategic long-lived Player 1, there exists a (public) strategy that only conditions on $P_\sigma(\omega | z_{[0,t-1]}^2) \in \Delta(\Omega)$ and t which yields the strategic long-lived Player 1 a weakly better payoff against the given sequence of strategies of Player 2s.*

⁹We note that even though Assumption 3.1 is a natural one, it implies an equilibrium selection by ruling out Perfect Bayesian Equilibria like the following one: Suppose the stage game is a *coordination* game with multiple equilibria. Suppose further that public signals observed by Player 2s happen to be *uninformative* about Player 1's actions such that in every period each z^2 can be observed with exactly the same positive probability independent of Player 1's action. In this case Player 2's beliefs about Player 1's type do not change over time. If Player 2's prior belief about Player 1 being of normal type is large enough the following grim-trigger type of strategies constitute a Perfect Bayesian Equilibrium: Fix an arbitrary z^{2*} , as long as $z_t^2 \neq z^{2*}$ Player 1 and Player 2 play according to a particular stage game equilibria. Starting from the first time period t^* such that the public signal $z_{t^*}^2 = z^{2*}$, Player 1 and Player 2 plays according to another particular stage game equilibrium forever. As one can see, Player 2s' strategy here is measurable with respect to the sigma-algebra over the histories they observe but not on their posterior beliefs. This is a limitation of Lemma 3.2 due to Assumption 3.1.

Proof. See Appendix.

3.2 Controlled Markov Chain Construction

The proof of Lemma 3.2 reveals the construction of a controlled Markov chain (or a Markov Decision Process); we refer the reader to the Appendix, Section A.9, for a brief review of this topic. Building on this proof, we will explicitly construct the dynamic programming problem as a controlled Markov chain optimization problem (that is, a *Markov Decision Process*). Under Assumption 3.1, given any sequence of strategies of Bayesian rational Player 2s, the solution to this optimization problem characterizes the equilibrium behavior of the strategic long-lived player in an associated Markov Perfect Equilibrium.

The state space, the action set, the transition kernel, and the per-stage reward function are given as follows:

- **The state space** is $\Delta(\Omega)$; $\mu_t \in \Delta(\Omega)$ is often called the *belief-state*. We endow this space with the weak convergence topology, and we note that since Ω is finite, the set of probability measures on Ω is a compact space.
- **The action set** is the set of all maps $\Gamma^1 := \{\gamma^1 : \Omega \rightarrow \mathbb{A}^1\}$. We note that since the commitment type policies are given a priori, one could also regard the action set to be the set \mathbb{A}^1 itself.¹⁰
- **The transition kernel** is given by $P : \Delta(\Omega) \times \Gamma^1 \rightarrow \mathcal{B}(\Delta(\Omega))$ ¹¹ so that for all $B \in \mathcal{B}(\Delta(\Omega))$:

$$\begin{aligned}
& P\left(P_\sigma(\omega|z_{[0,t-1]}^2) \in B \middle| P_\sigma(\omega|z_{[0,s-1]}^2), \gamma_s^1, s \leq t-1\right) \\
&= P\left(\left\{ \frac{\sum_{a_{t-1}^1} P_\sigma(z_{t-1}^2|a_{t-1}^1)P_\sigma(a_{t-1}^1|\omega, z_{[0,t-2]}^2)P_\sigma(\omega|z_{[0,t-2]}^2)}{\sum_{a_{t-1}^1, \omega} P_\sigma(z_{t-1}^2|a_{t-1}^1)P_\sigma(a_{t-1}^1|\omega, z_{[0,t-2]}^2)P_\sigma(\omega|z_{[0,t-2]}^2)} \right\} \in B \right. \\
&\quad \left. \middle| P_\sigma(\omega|z_{[0,s-1]}^2), \gamma_s^1, s \leq t-1\right) \\
&= P\left(\left\{ \frac{\sum_{a_{t-1}^1} P_\sigma(z_{t-1}^2|a_{t-1}^1)P_\sigma(a_{t-1}^1|\omega, z_{[0,t-2]}^2)P_\sigma(\omega|z_{[0,t-2]}^2)}{\sum_{a_{t-1}^1, \omega} P_\sigma(z_{t-1}^2|a_{t-1}^1)P_\sigma(a_{t-1}^1|\omega, z_{[0,t-2]}^2)P_\sigma(\omega|z_{[0,t-2]}^2)} \right\} \in B \right. \\
&\quad \left. \middle| P_\sigma(\omega|z_{[0,t-2]}^2), \gamma_{t-1}^1\right) \tag{4}
\end{aligned}$$

In the above derivation, we use the fact that the term $P_\sigma(a_{t-1}^1|\omega, z_{[0,t-2]}^2)$ is uniquely identified by $P_\sigma(\omega|z_{[0,t-2]}^2)$ and γ_{t-1}^1 . Here, γ_{t-1}^1 is the *control action*.

□

¹⁰We note that randomized strategies may also be considered by adding a randomization variable.

¹¹ $\mathcal{B}(\Delta(\Omega))$ is the set of all Borel sets on $\Delta(\Omega)$

- **The per-stage reward function**, given γ_t^2 , is $U(\mu_t, \gamma^1) : \Delta(\Omega) \times \Gamma^1 \rightarrow \mathbb{R}$ which is defined as follows

$$U(\mu_t, \gamma^1) := \sum_{\omega} P_{\sigma}(\omega | z_{[0,t-1]}^2) \sum_{\mathbb{A}^1} \left(1_{\{a_t^1 = \gamma^1(\omega)\}} u^1(a_t^1, \gamma_t^2(P_{\sigma}(a_t^1 | z_{[0,t-1]}^2), z_t^2)) \right) \quad (5)$$

where $\mu_t = P_{\sigma}(\omega | z_{[0,t-1]}^2)$. Here, γ_t^2 is a given measurable function of the posterior $P_{\sigma}(a_t^1 | z_{[0,t]}^2)$. We note again that for each Bayesian rational short-lived Player 2 we have

$$\gamma_t^2(P_{\sigma}(a_t^1 | z_{[0,t-1]}^2), z_t^2) \in \arg \max \left(\sum_{a^1} P_{\sigma}(a_t^1 | z_{[0,t]}^2) u^2(a^1, a^2) \right).$$

Remark 3.1. If $\arg \max(\sum_{a^1} P_{\sigma}(a_t^1 | z_{[0,t]}^2) u^2(a^1, a^2))$ is not unique, Player 2 may also randomize by picking randomly an action a^2 according to a fixed probability measure with support contained in the set $\arg \max(\sum_{a^1} P_{\sigma}(a_t^1 | z_{[0,t]}^2) u^2(a^1, a^2))$. The important requirement here is that each Player 2 plays the stage game only once, hence does not take into consideration future interactions.

We note here that Lemma 3.2 implies that in the finitely repeated setup, under Assumption 3.1, when Player 2s are Bayesian rational, the long-lived strategic Player 1 can depend his strategy only on Player 2s' posterior belief and time without any loss in payoff performance.

Consider now any Perfect Bayesian Equilibrium where the strategic long-lived Player 1 plays a private strategy, since the strategic long-lived Player 1 cannot have a profitable deviation, the public strategy identified in Lemma 3.2 must also give him the same payoff against the given sequence of strategies of Player 2s. Therefore, in the finitely repeated setup, under Assumption 3.1, any **Perfect Bayesian Equilibrium** payoff of the normal type Player 1, is also a **Perfect Public Equilibrium** payoff.¹²

Lemma 3.1 and Lemma 3.2 above have a coding theoretic flavor: The classic works by Witsenhausen [55] and Walrand and Varaiya [54], are of particular relevance; Teneketzis [53] extended these approaches to the more general setting of non-feedback communication and Yüksel [56] and Yüksel and Başar [57] extended these results to more general state spaces (including \mathbb{R}^d). Extension to infinite horizon stages have been studied in Linder and Yüksel [41]. In particular, Lemma 3.1 can be viewed as a generalization of Witsenhausen [55]. On the other hand, Lemma 3.2 can be viewed as a generalization of Walrand and Varaiya [54] and Linder and Yüksel [41]. The proofs build on the unified approach in Yüksel [56]. However, these results are different from the above contributions due to the fact that the utility functions do not depend explicitly on the type of Player 1, but depend explicitly on the actions a_t^1 and that these actions are not available to Player 2 unlike the setup in Yüksel [56]. Furthermore, we consider the infinitely repeated setup in the following.

3.3 Infinite Horizon and Equilibrium Strategies

We proceed with Lemma 3.3 which is the extension of Lemma 3.2 to the **the infinitely repeated setup**. Lemma 3.3 will be the key result which gives us a similar controlled

¹²A Perfect Public Equilibrium is a Perfect Bayesian Equilibrium where each player uses a public strategy, i.e., a strategy that only depends on the information which is available to both players.

Markov chain construction for the infinitely repeated game, hence a payoff-equivalent stationary Markov Perfect Equilibrium for each Perfect Bayesian Equilibrium which satisfies the following measurability / stationarity assumption for Player 2's strategies.

Assumption 3.2. *Each Player 2's strategy is a measurable function of $P_\sigma(a_t^1 | z_{[0,t]}^2)$ (that is, player 2s are stationary).*

Lemma 3.3. *In the infinitely repeated game, under Assumption 3.2, given any arbitrary sequence of strategies of Bayesian rational short-lived Player 2s, for any (private) strategy of the strategic long-lived Player 1, there exists a (public) strategy that only conditions on $P_\sigma(\omega | z_{[0,t-1]}^2) \in \Delta(\Omega)$ and t which yields the strategic long-lived Player 1 a weakly better payoff against the given sequence of strategies of Player 2s.*

Furthermore, the strategic long-lived Player 1's optimal stationary strategy against this given sequence of strategies of Player 2s can be characterized by solving an infinite horizon discounted dynamic programming problem.

Proof. See Appendix.

Therefore, in the infinitely repeated setup as well, under Assumption 3.1, any private strategy of the normal type Player 1 can be replaced, without any loss in payoff performance, with a public strategy which only depends on $P_\sigma(\omega | z_{[0,t-1]}^2)$ and t . Hence, for any **Perfect Bayesian Equilibrium** there exists a **Perfect Public Equilibrium** which is payoff-equivalent for the strategic long-lived Player 1 in the infinitely repeated game as well.

Furthermore, since there is a stationary optimal public strategy for the strategic long-lived Player 1 against any given sequence of strategies of Bayesian rational Player 2s, under Assumption 3.2, any payoff the strategic long-lived Player 1 obtains in a **Perfect Bayesian Perfect Equilibrium**, he can also obtain in a **Markov Perfect Equilibrium**.¹³ We state this result as our first main result, Theorem 3.1, below:

Theorem 3.1. *In the infinitely repeated game, under Assumption 3.2, the set of Perfect Bayesian Equilibrium payoffs of the strategic long-lived Player 1 is equal to the set of Markov Perfect Equilibrium payoffs.*

Proof. Markov Perfect Equilibrium payoff set is a subset of Perfect Bayesian Equilibrium payoff set. Hence, it is enough to show that for each Perfect Bayesian Equilibrium there exists a properly defined Markov Perfect Equilibrium which is payoff equivalent for the strategic long-lived Player 1. This follows from Lemma 3.3. \square

4 Undiscounted Average Payoff Case and An Upper Payoff Bound for the Arbitrarily Patient Long-lived Player

We next analyze the setup where the strategic long-lived Player 1 were to maximize his **undiscounted** average payoff instead of his discounted average payoff. Not only we identify

¹³A Markov Perfect Equilibrium is a Perfect Bayesian equilibrium where there is a payoff-relevant state space and both players are playing Markov strategies which only depends on the state variable.

an optimal strategy for the strategic long-lived Player 1 in this setup, but also we establish an **upper payoff bound** for the arbitrarily patient strategic long-lived Player 1 in the **standard discounted average payoff case** – through an Abelian inequality. To the best of our knowledge, such a technique does not seem to be considered in the reputations literature before.

The only difference from our original setup is that the strategic long-lived Player 1 now wishes to maximize

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \mathbb{E}_{\sigma^1, \sigma^2}^{\mu_0} \left[\sum_{t=0}^{N-1} u^1(a_t^1, a_t^2) \right].$$

Therefore, in any Perfect Bayesian Equilibrium, same as before, the short-lived (Bayesian rational) Player 2s will continue to be best replying to their updated beliefs. On the other hand, the strategic long-lived Player 1 will be playing a strategy which maximizes his undiscounted average payoff given that each Player 2 will be best replying to their updated beliefs.

The main problem in analyzing the undiscounted average payoff setup is that most of the structural coding/signaling results that we have for finite horizon or infinite horizon discounted optimal control problems do not generalize for the undiscounted case, see Linder and Yüksel [41]. Therefore, we will arrive at the following results using an indirect approach which is based on more intricate arguments.

Observe that $\{\mu_t(\bar{\omega}) = \mathbb{E}[1_{\omega=\bar{\omega}} | z_{[0,t]}^2]\}$, for every fixed $\bar{\omega}$, is a bounded martingale sequence adapted to the information at Player 2, and as a result as $t \rightarrow \infty$, by the submartingale convergence theorem [7] there exists $\bar{\mu}$ such that $\mu_t \rightarrow \bar{\mu}$ almost surely.

Let us re-visit the discounted average payoff setup: Let $\bar{\mu}$ be an **invariant posterior**, that is, a limit of the μ_t process which exists by the discussion with regard to the submartingale convergence theorem. Equation (12) leads to the following fixed point equation:

$$V^1(\omega, \mu) = \max_{a^1 = \gamma_t^1(\mu, \omega)} (\mathbb{E}[u^1(a_t^1, \gamma^2(\mu)) + \delta \mathbb{E}[V^1[(\omega, \mu)]])$$

Therefore,

$$V^1(\omega, \mu) = \frac{1}{1 - \delta} \max_{\gamma_t^1} \mathbb{E}[u^1(a_t^1, a_t^2(\mu))],$$

and since the solution is asymptotically stationary, the optimal strategy of the strategic long-lived Player 1 when $\mu_0 = \mu$ has to be a Stackelberg solution for a Bayesian game with prior μ ; thus, *a Perfect Bayesian Equilibrium strategy for the strategic long-lived Player 1 has to be mimicking the stage game Stackelberg type forever.*

Thus, every optimal strategy should be such that if Player 2's belief has converged, then the equilibrium behaviour must be of Stackelberg type. Note also that, by the analysis in the previous section, Player 2 behaves as if his strategy is optimal once his opinion is within a neighbourhood of the limit belief. Once Player 2s start best replying to the limit belief, Player 1's strategy becomes the Stackelberg action which is maximized according to the limit belief of Player 2s.

We state the following identifiability assumption.

Assumption 4.1. *Uniformly over all stationary optimal (for some discount parameter) strategies $\tilde{\sigma}^1, \tilde{\sigma}^2$,*

$$\lim_{\delta \rightarrow 1} \sup_{\tilde{\sigma}^1, \tilde{\sigma}^2} \left| \mathbb{E}_{\tilde{\sigma}^1, \tilde{\sigma}^2} (1 - \delta) \left[\sum_{t=0}^{\infty} \delta^t u^1(a_t^1, a_t^2) \right] - \limsup_{N \rightarrow \infty} \frac{1}{N} \mathbb{E}_{\tilde{\sigma}^1, \tilde{\sigma}^2} \left[\sum_{t=0}^{N-1} u^1(a_t^1, a_t^2) \right] \right| = 0 \quad (6)$$

Assumption 4.1 may seem to be a strict assumption at first look. Below, we explain why this does not happen to be the case. We first note that a sufficient condition for Assumption 4.1 is the following:

Assumption 4.2. *Whenever the strategic long-lived Player 1 adopts a stationary strategy, for any initial commitment prior, there exists a stopping time τ such that for $t \geq \tau$, Player 2s' posterior beliefs become so that his best response does not change (that is, his best-response to his beliefs leads to a constant action). Furthermore, $\mathbb{E}[\tau] < \infty$, uniformly over any stationary strategy σ^1 .*

Furthermore, Proposition 4.1 below shows that Assumption 4.1 is indeed implied by one of the most standard identifiability assumptions in the repeated games literature:

Proposition 4.1. *Consider the matrix A whose rows consist of the vectors:*

$$[P_{\sigma}(z_t^2 = k | a_t^1 = 1) \quad P_{\sigma}(z_t^2 = k | a_t^1 = 2) \quad \cdots \quad P_{\sigma}(z_t^2 = k | a_t^1 = |\mathbb{A}^1|)]$$

where $k \in \{1, 2, \dots, |\mathbb{Z}^2|\}$. If $\text{rank}(A) = |\mathbb{A}^1|$, then Assumption 4.1 holds.

Proof. See Appendix.

As was mentioned above, the sufficient condition described in Proposition 4.1 is a standard identifiability assumption, sometimes referred as the full-rank monitoring assumption in the reputations literature, see for example Assumption 2 of Cripps, Mailath, and Samuelson [14]. Therefore, Assumption 4.1 is a relatively reasonable identifiability assumption which is weaker than one of the most standard identifiability assumptions in the reputations literature. Under Assumption 4.1, we establish that mimicking a Stackelberg commitment type forever is an optimal strategy for the strategic long-lived Player 1 in the undiscounted average payoff setup:

Theorem 4.1. *In the undiscounted average payoff setup, under Assumption 4.1, an optimal strategy for the strategic long-lived Player 1 in the infinitely repeated game is the stationary strategy “mimicking the Stackelberg commitment type forever.”*

Proof. See Appendix.

As an implication of Theorem 4.1, we next state the aforementioned upper bound for Perfect Bayesian Equilibrium payoffs of the arbitrarily patient strategic long-lived Player 1 in the standard discounted average payoff setup as Theorem 4.2.

Theorem 4.2. *Under Assumption 4.1, $\limsup_{\delta \rightarrow 1} V_\delta^1(\omega, \mu^0) \leq \max_{\alpha_1 \in \Delta(A_1), \alpha_2 \in BR(\alpha_1)} u_1(\alpha_1, \alpha_2)$. That is, an upperbound for the value of the reputation for an arbitrarily patient strategic long-lived Player 1 in any Perfect Bayesian Equilibrium is his stage game Stackelberg equilibrium payoff.*

Proof of Theorem 4.2. Note the following Abelian theorem: Let a_n be a sequence of non-negative numbers and $\beta \in (0, 1)$. Then,

$$\begin{aligned} \liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{m=0}^{N-1} a_m &\leq \liminf_{\beta \uparrow 1} (1 - \beta) \sum_{m=0}^{\infty} \beta^m a_m \\ &\leq \limsup_{\beta \uparrow 1} (1 - \beta) \sum_{m=0}^{\infty} \beta^m a_m \leq \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{m=0}^{N-1} a_m \end{aligned} \quad (7)$$

Therefore, for any δ , an upper bound is obtained by the corresponding undiscounted average payoff problem. Since for every δ , an optimal strategy is stationary, and under the stationary strategy the average payoff converges to the one achieved by the case where the type of Player 1 is correctly identified by Player 2s, the result follows from Theorem 4.1. \square

Theorem 4.2 provides an upper bound on the value of reputation for the strategic long-lived Player 1 in the standard discounted average payoff setup. That is, in the standard discounted average payoff setup, an arbitrarily patient strategic long-lived Player 1 cannot do any better than his best Stackelberg payoff under reputational concerns as well. This upperbound coincides with those provided before by Fudenberg and Levine [26] and Gossner [30].

5 A Lower Payoff Bound on Reputation through Measure Concentration

We next identify a lower payoff bound for the value of reputation through an explicit measure concentration analysis. As mentioned before, it was Fudenberg and Levine [25], [26] who provided such a lower payoff bound for the first time. They constructed a lower bound for any equilibrium payoff of the strategic long-lived player by showing that Bayesian rational short-lived players can be surprised at most finitely many times when a strategic long-lived Player mimics a commitment type forever.

Gossner [30], on the other hand, used information theoretic ideas to obtain a more concise lower payoff bound: Using the chain rule property of the concept of relative entropy, Gossner [30] obtained a lower bound for any equilibrium payoff of the strategic long-lived player by showing that any equilibrium payoff of the strategic long-lived player is bounded from below (and above) by a function of the average discounted divergence between the prediction of the short-lived players conditional on the long-lived player's type and its marginal.

Our analysis below provides a sharper lower payoff bound for the value of reputation through a refined measure concentration analysis. To obtain this lower bound, as in Fudenberg and Levine [26] as well as Gossner [30], we let the strategic long-lived Player 1 mimic (forever) a commitment type, $\hat{\omega} = m$, to investigate the best responses of the short-lived

Player 2s. In any Perfect Bayesian Equilibrium, such a deviation, i.e. deviating to mimicking a particular commitment type forever, is always possible for the strategic-long lived Player 1.

Let $|\Omega| = M$ be the number of all possible types of the long-lived Player 1. With m being the type mimicked forever by Player 1, we will identify a function f below such that for any $\hat{\omega} \in \hat{\Omega}$ when criterion (8) below holds,

$$\frac{P_\sigma(\omega = m | z_{[0,t]}^2)}{P_\sigma(\omega = \hat{\omega} | z_{[0,t]}^2)} \geq f(M), \quad (8)$$

Player 2 of time t will act as if he knew the type of the long-lived Player 1 is m . This will follow from the fact that $\max_{a^2} \sum P_\sigma(\hat{\omega} | z_{[0,t]}^2) u^2(a^1, a^2)$ is continuous in $P_\sigma(\hat{\omega} | z_{[0,t]}^2)$ and that $P_\sigma(\hat{\omega} | z_{[0,t]}^2)$ concentrates around the true type under a mild informativeness condition on the observable variables.

Let

$$\begin{aligned} \tau_m &= \min\{T \geq 0 : \max_{a^2} \sum_{a^1} P_\sigma(a^1 | z_{[0,t]}^2) u^2(a^1, a^2) \\ &= \max_{a^2} \sum_{a^1} P_\sigma(a^1 | \omega = m) u^2(a^1, a^2) \quad \forall t \geq T\}, \end{aligned}$$

Intuitively, τ_m is the time when Player 2s start to behave as if the type of the long-lived Player 1 is m as far as their optimal strategies are concerned.

The following lemma provides an upper bound for τ_m regarding the aforementioned criterion (8).

Lemma 5.1. *Let $\epsilon > 0$ be such that for any $\bar{a}^1 \in \mathbb{A}^1$ and $\tilde{a}^2, \hat{a}^2 \in \mathbb{A}^2$*

$$|u^2(\bar{a}^1, \tilde{a}^2) - u^2(\bar{a}^1, \hat{a}^2)| \geq \frac{\epsilon}{1 - \epsilon} \left(\max_{a^1, a^2} |u^2(a^1, a^2)| \right)$$

If (8) holds at time t when $f(M) = \frac{(1-\epsilon)}{\epsilon} M$, then $\tau_m \leq t$.

Proof. See Appendix.

Lemma 5.1 implies that when criterion (8) holds to be true for $f(M) = \frac{(1-\epsilon)}{\epsilon} M$, at time t any Player 2 of time t and onwards will be best responding to the commitment type m . This can be interpreted as the long-lived Player **having a reputation to behave like type m** when criterion (8) is satisfied.

We next provide Theorem 5.1 which shows that as a stopping time, τ_m is dominated by a geometric random variable.

Theorem 5.1. *Suppose that $0 < \frac{P_\sigma(z^2 | \omega = m)}{P_\sigma(z^2 | \omega = \hat{\omega})} < \infty$ for all $\hat{\omega} \in \hat{\Omega}$ and $z^2 \in \mathbb{Z}^2$. For all $k \in \mathbb{N}$, $P_\sigma(\tau_m \geq k) \leq R \rho^k$ for some $\rho \in (0, 1)$ and $R \in \mathbb{R}$.*

Proof. See Appendix.

We are now ready to provide our lower bound for Perfect Bayesian Equilibrium payoffs of the strategic long-lived Player 1, for a fixed discount factor $\delta \in (0, 1)$.

Theorem 5.2. *A lower bound for the expected payoff of the strategic long-lived Player 1 in any Perfect Bayesian Equilibrium is given by $\max_{m \in \hat{\Omega}} L(m)$ where*

$$L(m) = \mathbb{E}_{\{\omega=m\}} \left[\sum_{k=1}^{\tau_m} \delta^k u^1(a_t^1, a_t^2) \right] + \mathbb{E}_{\{\omega=m\}} \left[\sum_{k=\tau_m+1}^{\infty} \delta^k \underline{u}_s^{1*}(m) \right]$$

where $\underline{u}_s^{1*}(m) := \min_{a^2 \in BR^2(m)} u^1(m, a^2)$ and $BR^2(m) := \arg \max_{a^2 \in \mathbb{A}^2} u^2(m, a^2)$.

Proof. It follows from Theorem 5.1 that the stopping time τ_m is dominated by a geometric random variable. Therefore the discounted average payoff can be lower bounded by the sum of the following two terms:

$$\mathbb{E}_{\{\omega=m\}} \left[\sum_{k=1}^{\tau_m} \delta^k u^1(a_t^1, a_t^2) \right] + \mathbb{E}_{\{\omega=m\}} \left[\sum_{k=\tau_m+1}^{\infty} \delta^k \underline{u}_s^{1*}(m) \right]$$

where $\underline{u}_s^{1*}(m) := \min_{a^2 \in BR^2(m)} u^1(m, a^2)$ and $BR^2(m) := \arg \max_{a^2 \in \mathbb{A}^2} u^2(m, a^2)$. Since a deviation to mimicking any of the commitment types forever is available to the strategic long-lived Player 1 in any Perfect Bayesian Equilibrium, taking the maximum of the lower bound above for all commitment types gives the desired result. \square

Observe that when m is a Stackelberg type, i.e., a commitment type who is committed to play the stage game Stackelberg action $\arg \max_{\alpha_1 \in \Delta(A_1)} u_1(\alpha_1, BR^2(\alpha_1))$ for which Player 2s have a unique best reply then $\underline{u}_s^{1*}(m) = \max_{\alpha_1 \in \Delta(A_1), \alpha_2 \in BR(\alpha_1)} u_1(\alpha_1, \alpha_2)$ becomes the stage game Stackelberg payoff.

We next turn to the case of the arbitrarily patient strategic long-lived Player 1. That is, what happens when $\delta \rightarrow 1$.

Theorem 5.3.

$$\lim_{\delta \rightarrow 1} L(m) \geq \underline{u}_s^{1*}(m)$$

Proof of Theorem 5.3. Their proof follows from Theorem 5.2 by taking the limit $\delta \rightarrow 1$. Since until time τ_m , we can bound the payoff to strategic long-lived Player 1 below by the worst possible payoff, and after τ_m the strategic long-lived Player 1 guarantees the associated Stackelberg payoff, we obtain

$$\lim_{\delta \rightarrow 1} L(m) \geq \lim_{\delta \rightarrow 1} \mathbb{E}[1 - \delta^{\tau_m}] \min_{a^1, a^2} u^1(a^1, a^2) + \lim_{\delta \rightarrow 1} \mathbb{E}[\delta^{\tau_m} \underline{u}_s^{1*}(m)] = \underline{u}_s^{1*}(m).$$

That $\lim_{\delta \rightarrow 1} \mathbb{E}[\delta^{\tau_m} - 1] = 0$ and $\lim_{\delta \rightarrow 1} \mathbb{E}[\delta^{\tau_m}] = 1$ follow from the dominated convergence theorem and that τ_m is finite with probability 1. \square

Theorem 5.3 implies that if there exists a Stackelberg commitment type as defined before, an arbitrarily patient strategic long-lived Player 1 can guarantee himself a payoff arbitrarily close to the associated Stackelberg payoff in every Perfect Bayesian Equilibrium. This means that, as was mentioned before, the lower payoff bound that we provided in Theorem 5.2 coincides in the limit as $\delta \rightarrow 1$ with those of Fudenberg and Levine [26] and Gossner [30].

6 Continuity of payoff values

Next, we consider the continuity of the payoff values of the strategic long-lived Player 1 in the prior beliefs of Player 2s for any Markov Perfect Equilibrium obtained through the aforementioned dynamic programming.

Lemma 6.1. *The transition kernel of the aforementioned Markov chain is weakly continuous in the (belief) state and action.*

Proof. See Appendix.

Next, we note that any strategy for some Player 2 of time t which chooses

$$\arg \max_{a^1} \left(\sum_{a^1} P_\sigma(a_t^1 | z_{[0,t]}^2) u^2(a^1, a^2) \right)$$

in a measurable fashion does not have to be continuous in the conditional probability $\kappa(\cdot) = P_\sigma(a_t^1 = \cdot | z_{[0,t]}^2)$, since such a strategy partitions (or quantizes) the set of probability measures on \mathbb{A}^1 . The set of κ for which this discontinuity holds is a subset of the set of probability measures $\mathcal{B}_e = \cup_{k,m \in \mathbb{A}^2} \mathcal{B}^{k,m}$, where for some pair $k, m \in \mathbb{A}^2$, the set $\mathcal{B}^{k,m}$ is defined as

$$\mathcal{B}^{k,m} = \sum_{a^1 \in \mathbb{A}^1} \kappa(a^1) u^2(a^1, k) = \sum_{a^1 \in \mathbb{A}^1} \kappa(a^1) u^2(a^1, m).$$

These are the sets of probability measures where Player 2 is indifferent between two actions.

However, even though the strategy of Player 2s may not be continuous in the conditional probability κ , the *per-stage reward function*, $U(\mu_t, \gamma^1)$ may be continuous in μ_t , which we list as Assumption 6.1 below.

Assumption 6.1. *The per-stage reward function, $U(\mu_t, \gamma^1)$, is continuous in μ_t .*

Assumption 6.1 is a rather technical assumption which might be demanding. However, this is an essential condition to be able to have continuity in the pay-off functions. In the following remark we list two sufficient conditions for Assumption 6.1.

Remark 6.1. Two sufficient conditions for the continuity of the per-stage reward function $U(\mu_t, \gamma^1)$ in μ_t are as follows.

- (i) If the stage game payoff functions for the players are identical or are aligned as in a potential game, using the fact that if f is a continuous function, for a compact \mathbb{U} , the function $\min_{u \in \mathbb{U}} f(x, u)$ is continuous in x , continuity can be established; see e.g. the proof of Theorem 4 in Linder and Yüksel [41] for a related discussion in the context of identical payoff functions. In our setup, however, when the payoff functions for Players 1 and 2s are not aligned, this is not guaranteed and counterexamples can be constructed, see Cripps and Faingold [13].
- (ii) Another sufficient condition for Assumption 6.1 to hold is that the probability measure $P_\sigma(P_\sigma(a_t^1 = \cdot | z_{[0,t]}^2) \in \mathcal{B}_e) = 0$ for all t values. In particular, if players 2 always have a unique best response so that the set of discontinuity, \mathcal{B}_e , is never visited (with probability 1), then Assumption 6.1 holds as well.

The continuity of the transition kernel and per-stage reward function together with the compactness of the action space leads to the desired continuity result.

Theorem 6.1. *Under Assumption 6.1, the value function V_t^1 of the dynamic program given in (12) is continuous in μ_t for all $t \geq 0$.*

Proof of Theorem 6.1. Given Lemma 6.1 and Assumption 6.1, the proof follows from an inductive argument and the measurable selection hypothesis (see Theorem A.2 in Appendix A.9). In this case, the discounted optimality operator becomes a contraction mapping from the Banach space of continuous functions on $\Delta(\Omega)$ to itself, leading to a fixed point in this space. \square

Theorem 6.1 implies that, any Markov Perfect Equilibrium payoff of the strategic long-lived Player 1 obtained through the dynamic program in (12) is robust to small perturbations in the prior beliefs of Player 2s. This further implies that the following conjecture made by Cripps, Mailath, and Samuelson [14] is indeed true in our setup under Assumption 6.1: There exists a particular equilibrium in the complete information game and a bound such that for *any* commitment type prior (of Player 2s) less than this bound, there exists an equilibrium of the incomplete information game where the long-lived player’s payoff is arbitrarily close to his payoff from this particular equilibrium of the complete information game.¹⁴ This is also in line with the findings of Dalkıran [16] who uses the methods of Abreu, Pearce, and Stacchetti [1] to show a similar continuity result.

7 Conclusion

In this paper, we studied the reputations problem of an informed long-lived player who controls his reputation against a sequence of uninformed short-lived players by employing tools from stochastic control theory. The main assumption in our model was that the information of the short-lived players is nested in that of the long-lived player. Under this assumption:

- (i) We showed that, given mild assumptions, the set of Perfect Bayesian Equilibrium payoffs coincide with Markov Perfect Equilibrium payoffs in the standard discounted average payoff setup.
- (ii) A dynamic programming formulation was obtained for the computation of equilibrium strategies of the strategic long-lived player in the standard discounted average payoff setup.
- (iii) We also considered the undiscounted average payoff setup separately where we obtained an optimal equilibrium strategy of the strategic long-lived player.
- (iv) We then used this optimal strategy in the undiscounted average payoff setup as a tool to obtain an upper payoff bound for the arbitrarily patient long-lived player in the standard discounted average payoff setup.
- (v) By using measure concentration techniques, we obtained a refined lower payoff bound on the value of reputation in the standard discounted average payoff setup.

¹⁴This conjecture appears as a presumption of Theorem 3 in Cripps, Mailath, and Samuelson [14]. They write “We conjecture this hypothesis is redundant, given the other conditions of the theorem, but have not been able to prove it”.

(vi) Finally, under further assumptions, we established the continuity of the equilibrium payoffs, in the standard discounted average payoff setup.

Our findings contribute to the reputations literature by obtaining new results on the structure of equilibrium behavior in finite-horizon, infinite-horizon and undiscounted settings, as well as continuity results in the prior probabilities, and improved upper and lower bounds on the value of reputations. In particular, we exhibited that a control theoretic formulation can be utilized to characterize the equilibrium behavior. It is our hope that the machinery we provide in this paper will open a new avenue for applied work studying reputations in different frameworks.

A Appendix

A.1 Proof of Lemma 3.1.

At time $t = T$, the payoff function can be written as follows, where γ_t^2 denotes a given fixed strategy for Player 2:

$$\mathbb{E}[u^1(a_t^1, \gamma_t^2(z_{[0,t]}^2)) | z_{[0,t-1]}^2] = \mathbb{E}[F(a_t^1, z_{[0,t-1]}^2, z_t^2) | z_{[0,t-1]}^2]$$

where, $F(a_t^1, z_{[0,t-1]}^2, z_t^2) = u^1(a_t^1, \gamma_t^2(z_{[0,t]}^2))$.

Now, by a stochastic realization argument (see Borkar [8]), we can write $z_t^2 = R(a_t^1, v_t)$ for some independent noise process v_t . As a result, the expected payoff conditioned on $z_{[0,t-1]}^2$ is equal to, by the smoothing property of conditional expectation, the following:

$$\mathbb{E} \left[\mathbb{E}[G(a_t^1, z_{[0,t-1]}^2, v_t) | \omega, a_t^1, z_{[0,t-1]}^2] \middle| z_{[0,t-1]}^2 \right],$$

for some G . Since v_t is independent of all the other variables at times $s \leq t$, it follows that there exists H so that $\mathbb{E}[G(a_t^1, z_{[0,t-1]}^2, v_t) | \omega, a_t^1, z_{[0,t-1]}^2] =: H(\omega, a_t^1, z_{[0,t-1]}^2)$. Note that when ω is a commitment type, a_t^1 is fixed quantity or a fixed random variable.

Now, we will apply Witsenhausen's two stage lemma [55], to show that we can obtain a lower bound for the double expectation by picking a_t^1 as a result of a measurable function of $\omega, z_{[0,t-1]}^2$. Thus, we will find a strategy which only uses $(\omega, z_{[0,t-1]}^2)$ which performs as well as one which uses the entire memory available at Player 1. To make this precise, let us fix γ_t^2 and define for every $k \in \mathbb{A}^1$:

$$\beta_k := \left\{ \omega, z_{[0,t-1]}^2 : G(\omega, z_{[0,t-1]}^2, k) \leq G(\omega, z_{[0,t-1]}^2, q), \forall q \neq k \right\}.$$

Such a construction covers the domain set consisting of $(x_t, q_{[0,t-1]})$ but possibly with overlaps. It covers the elements in $\Omega \times \prod_{t=0}^{T-1} \mathbb{Z}^2$, since for every element in this product set, there is a maximizing $k \in \mathbb{A}^1$. To avoid the overlap, define a function $\gamma_t^{*,1}$ as:

$$q_t = \gamma_t^{*,1}(\omega, z_{[0,t-1]}^2) = k, \quad \text{if } (\omega, z_{[0,t-1]}^2) \in \beta_k \setminus \cup_{i=1}^{k-1} \beta_i,$$

with $\beta_0 = \emptyset$. The new strategy performs at least as well as the original strategy even though it has a restricted structure.

The same discussion applies for earlier time stages as we discuss below. We iteratively proceed to study the other time stages. For a three-stage problem, the payoff at time $t = 2$ can be written as:

$$\mathbb{E} \left[u^1(a_2^1, \gamma_2^2(z_1^2, z_2^2)) + \mathbb{E}[u^1(\gamma_3^{*,1}(\omega, z_{[1,2]}^2), \gamma_3^2 \left(z_1^2, z_2^2, R(\gamma_3^{*,1}(\omega, z_{[1,2]}^2), v_3) \right) | \omega, z_1^2, z_2^2] \middle| z_1^2 \right]$$

The expression inside the expectation is equal to for some measurable F_2 , $F_2(\omega, a_2^1, z_1^2, z_2^2)$. Now, once again expressing $z_2^2 = R(a_2^1, v_2)$, by a similar argument as above, a strategy at time 2 which uses ω and z_1^2 and which performs at least as good as the original strategy can be constructed. By similar arguments, a strategy at time t , $1 \leq t \leq T$ only uses $(\omega, z_{[1,t-1]}^2)$ can be constructed. The strategy at time $t = 0$ uses ω . \square

A.2 Proof of Lemma 3.2.

The proof follows from a similar argument as that for Lemma 3.1, except that the information at Player 2 is replaced by the sufficient statistic that Player 2 uses: his posterior information. At time $t = T - 1$, an optimal Player 2 will use $P_\sigma(a_t^1|z_{[0,t]}^2)$ as a sufficient statistic for an optimal decision. Let us fix a strategy for Player 2 at time t , γ_t^2 which only uses the posterior $P_\sigma(a_t^1|z_{[0,t]}^2)$ as its sufficient statistic. Let us further note that:

$$\begin{aligned} P_\sigma(a_t^1|z_{[0,t]}^2) &= \frac{P_\sigma(z_t^2, a_t^1|z_{[0,t-1]}^2)}{\sum_{a_t^1} P_\sigma(z_t^2, a_t^1|z_{[0,t-1]}^2)} \\ &= \frac{\sum_\omega P_\sigma(z_t^2|a_t^1)P_\sigma(a_t^1|\omega, z_{[0,t-1]}^2)P_\sigma(\omega|z_{[0,t-1]}^2)}{\sum_\omega \sum_{a_t^1} P_\sigma(z_t^2|a_t^1)P_\sigma(a_t^1|\omega, z_{[0,t-1]}^2)P_\sigma(\omega|z_{[0,t-1]}^2)} \end{aligned} \quad (9)$$

The term $P_\sigma(a_t^1|\omega, z_{[0,t-1]}^2)$ is determined by the strategy of Player 1 (this follows from Lemma 3.1), γ_t^1 .

As in Yüksel and Başar [57], this implies that the payoff at the last stage conditioned on $z_{[0,t-1]}^2$ is given by

$$\mathbb{E} \left[u^1 \left(a_t^1, \gamma_t^2(P_\sigma(a_t^1 = \cdot | z_{[0,t]}^2)) \right) | z_{[0,t-1]}^2 \right] = \mathbb{E} \left[F \left(a_t^1, \gamma_t^1, P_\sigma(\omega = \cdot | z_{[0,t-1]}^2) \right) | z_{[0,t-1]}^2 \right]$$

where, as earlier, we use the fact that z_t^2 is conditionally independent of all the other variables at times $s \leq t$ given a_t^1 . Let $\gamma_t^{1, z_{[0,t-1]}^2}$ denote the strategy of Player 1. The above state is then equivalent to, by the smoothing property of conditional expectation, the following:

$$\begin{aligned} &\mathbb{E} \left[\mathbb{E} \left[F \left(a_t^1, \gamma_t^1, P_\sigma(\omega = \cdot | z_{[0,t-1]}^2) \right) | \omega, \gamma_t^{1, z_{[0,t-1]}^2}, P_\sigma(\omega = \cdot | z_{[0,t-1]}^2), z_{[0,t-1]}^2 \right] | z_{[0,t-1]}^2 \right] \\ &= \mathbb{E} \left[\mathbb{E} \left[F \left(a_t^1, \gamma_t^1, P_\sigma(\omega = \cdot | z_{[0,t-1]}^2) \right) | \omega, \gamma^{1, z_{[0,t-1]}^2}, P_\sigma(\omega = \cdot | z_{[0,t-1]}^2) \right] | z_{[0,t-1]}^2 \right] \end{aligned} \quad (10)$$

The second line follows since once one picks the strategy $\gamma^{1, z_{[0,t-1]}^2}$, the dependence on $z_{[0,t-1]}^2$ is redundant given $P_\sigma(\omega = \cdot | z_{[0,t-1]}^2)$.

Now, one can construct an equivalence class among the past $z_{[0,t-1]}^2$ sequences which induce the same $\mu_t(\cdot) = P_\sigma(\omega \in \cdot | z_{[0,t-1]}^2)$, and can replace the strategy in this class with one, which induces a higher payoff among the finitely many elements in each class for the final time stage. An optimal output thus may be generated using μ_t and ω and t , by extending Witsenhausen's argument used earlier in the proof of Lemma 3.1 for the terminal time stage. Since there are only finitely many past sequences and finitely many μ_t , this leads to a (Borel measurable) selection of ω for every μ_t , leading to a measurable strategy in μ_t, ω . Hence, the final stage payoff can be expressed as $F_t(\mu_t)$ for some F_t , without any performance loss.

The same argument applies for all time stages. To show this, we will apply induction as in Yüksel [56]. At time $t = T - 1$, the sufficient statistic both for the immediate payoff,

and the *continuation payoff* is $P_\sigma(\omega|z_{[0,t-1]}^2)$, and thus for the payoff impacting the time stage $t = T$, as a result of the optimality result for γ_T^1 . To show that the separation result generalizes to all time stages, it suffices to prove that $\{(\mu_t, \gamma_t^1)\}$ has a controlled Markov chain form, if the players use the structure above.

Now, for $t \geq 1$, for all $B \in \mathcal{B}(\Delta(\Omega))$:

$$\begin{aligned}
& P\left(P_\sigma(\omega|z_{[0,t-1]}^2) \in B \middle| P_\sigma(\omega|z_{[0,s-1]}^2), \gamma_s^1, s \leq t-1\right) \\
&= P\left(\left\{ \frac{\sum_{a_{t-1}^1} P_\sigma(z_{t-1}^2|a_{t-1}^1) P_\sigma(a_{t-1}^1|\omega, z_{[0,t-2]}^2) P_\sigma(\omega|z_{[0,t-2]}^2)}{\sum_{a_{t-1}^1, \omega} P_\sigma(z_{t-1}^2|a_{t-1}^1) P_\sigma(a_{t-1}^1|\omega, z_{[0,t-2]}^2) P_\sigma(\omega|z_{[0,t-2]}^2)} \right\} \in B \right. \\
&\quad \left. \middle| P_\sigma(\omega|z_{[0,s-1]}^2), \gamma_s^1, s \leq t-1\right) \\
&= P\left(\left\{ \frac{\sum_{a_{t-1}^1} P_\sigma(z_{t-1}^2|a_{t-1}^1) P_\sigma(a_{t-1}^1|\omega, z_{[0,t-2]}^2) P_\sigma(\omega|z_{[0,t-2]}^2)}{\sum_{a_{t-1}^1, \omega} P_\sigma(z_{t-1}^2|a_{t-1}^1) P_\sigma(a_{t-1}^1|\omega, z_{[0,t-2]}^2) P_\sigma(\omega|z_{[0,t-2]}^2)} \right\} \in B \right. \\
&\quad \left. \middle| P_\sigma(\omega|z_{[0,s-1]}^2), \gamma_s^1, s = t-1\right) \quad (11)
\end{aligned}$$

In the above derivation, we use the fact that the term $P_\sigma(a_{t-1}^1|\omega, z_{[0,t-2]}^2)$ is uniquely identified by $P_\sigma(\omega|z_{[0,t-2]}^2)$ and γ_{t-1}^1 . \square

A.3 Proof of Lemma 3.3.

First, going from a finite horizon to an infinite horizon follows from a change of order of limit and infimum as we discuss in the following. Observe that for any strategy $\{\gamma_t^1\}$ and any $T \in \mathbb{N}$:

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \delta^t u^1(a_t^1, a_t^2)\right] \geq \inf_{\{\gamma_t^1\}} \mathbb{E}\left[\sum_{t=0}^{T-1} \delta^t u^1(a_t^1, a_t^2)\right]$$

and thus

$$\lim_{T \rightarrow \infty} \mathbb{E}\left[\sum_{t=0}^{T-1} \delta^t u^1(a_t^1, a_t^2)\right] \geq \limsup_{T \rightarrow \infty} \inf_{\{\gamma_t^1\}} \mathbb{E}\left[\sum_{t=0}^{T-1} \delta^t u^1(a_t^1, a_t^2)\right]$$

Since the above holds for an arbitrary strategy, it follows then that

$$\inf_{\{\gamma_t^1\}} \lim_{T \rightarrow \infty} \mathbb{E}\left[\sum_{t=0}^{T-1} \delta^t u^1(a_t^1, a_t^2)\right] \geq \limsup_{T \rightarrow \infty} \inf_{\{\gamma_t^1\}} \mathbb{E}\left[\sum_{t=0}^{T-1} \delta^t u^1(a_t^1, a_t^2)\right]$$

On the other hand, due to the discounted nature of the problem, the right hand side can be studied through the dynamic programming (Bellman) iteration algorithms: The following dynamic program holds: Let $\mu_t(w) = P_\sigma(\omega = w|z_{[0,t-1]}^2)$.

$$V^1(\omega, \mu_t) = \max_{\gamma_t^1} \left(\mathbb{E}\left[u^1(a_t^1, a_t^2) + \delta \mathbb{E}[V^1(\omega, \mu_{t+1}) | \mu_t, \gamma_t^1]\right] \right)$$

$$=: \mathbb{T}(V^1)(\omega, \mu_t) \quad (12)$$

where \mathbb{T} is an operator defined by:

$$\mathbb{T}(f)(\omega, \mu_t) = \max_{\gamma_t^1} \left(\mathbb{E} \left[u^1(a_t^1, a_t^2) + \delta \mathbb{E}[f(\omega, \mu_{t+1}) | \mu_t, \gamma_t^1] \right] \right)$$

A value iteration sequence with $V_0^1 = 0$ and $V_{t+1} = \mathbb{T}(V_t)$ leads to a stationary solution. This is an infinite horizon discounted payoff optimal dynamic programming equation with a compact state (belie) space and a finite action spaces (where the strategy is now the *action* γ_t^1). Since the action set is finite in our formulation, By Theorem A.3 (see Appendix A.9), it follows that there is a stationary solution as $t \rightarrow \infty$.

Thus, the sequence of maximizations $\sup_{\gamma^1} \mathbb{E}[\sum_{t=0}^{T-1} \delta^t u^1(a_t^1, a_t^2)]$ leads to a stationary solution as $T \rightarrow \infty$, and this sequence of policies admit the structure stated in the statement of the theorem. As a result, we can state that such strategies are optimal also for the infinite horizon setup, and the dependence on t is eliminated. \square

A.4 Proof of Proposition 4.1.

Following Gossner [30], we know that the conditional relative entropy

$$\mathbb{E} \left[D \left(P_\sigma(z_t^2 \in \cdot | h_t^2, \omega) || P_\sigma(z_t^2 \in \cdot | h_t^2) \right) \right] \rightarrow 0$$

and Pinsker's inequality that convergence in total variation is implied by convergence in relative entropy; it follows that for every $z \in \mathbb{Z}^2$

$$\mathbb{E}[|P_\sigma(z_t^2 = z | h_t^2) - P_\sigma(z_t^2 = z | h_t^2, \omega)|] \rightarrow 0 \quad (13)$$

But,

$$P_\sigma(z_t^2 = z | h_t^2) = \sum_{a^1} P_\sigma(z_t^2 = z | a_t^1 = a^1) P_\sigma(a_t^1 = a^1 | h_t^2)$$

Thus, all we need to ensure is that Player 2's belief on $P_\sigma(a_t^1 | h_t^2)$ is sufficiently close to a terminal value. Suppose that the conditions of the theorem holds, but $|P_\sigma(a_t^1 | h_t^2) - P_\sigma(a_t^1 | h_t^2, \omega)| > \delta$ for some subsequence of time values. If the rank of A is $|\mathbb{A}^1|$, then, $|P_\sigma(a_t^1 | h_t^2) - P_\sigma(a_t^1 | h_t^2, \omega)| > \delta$ would imply that $|P_\sigma(z_t^2 | h_t^2) - P_\sigma(z_t^2 | h_t^2, \omega)| > \epsilon$ for some positive ϵ , which would be a contradiction (to see this, observe that the vector $P_\sigma(a_t^1 | h_t^2) - P_\sigma(a_t^1 | h_t^2, \omega)$ cannot be orthogonal to each of the rows of A , due to the rank condition). In particular, (14) implies the convergence of $P_\sigma(a_t^1 | h_t^2)$ to a limit. Furthermore, Jensen's inequality implies that

$$|\mathbb{E}[P_\sigma(z_t^2 = z | h_t^2) - P_\sigma(z_t^2 = z | h_t^2, \omega)]| \leq \mathbb{E}[|P_\sigma(z_t^2 = z | h_t^2) - P_\sigma(z_t^2 = z | h_t^2, \omega)|] \rightarrow 0 \quad (14)$$

and thus in finite expected time the deviation in the conditional probabilities will be less than a prescribed amount and Assumption 4.1 holds. \square

A.5 Proof of Theorem 4.1

Note the following *Abelian* inequalities (see, e.g., Lemma 5.3.1 in Hernandez-Lerma and Lasserre [31]): Let a_n be a sequence of non-negative numbers and $\beta \in (0, 1)$. Then,

$$\begin{aligned} \liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{m=0}^{N-1} a_m &\leq \liminf_{\beta \uparrow 1} (1 - \beta) \sum_{m=0}^{\infty} \beta^m a_m \\ &\leq \limsup_{\beta \uparrow 1} (1 - \beta) \sum_{m=0}^{\infty} \beta^m a_m \leq \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{m=0}^{N-1} a_m \end{aligned} \quad (15)$$

Thus, for every strategy pair σ^1, σ^2 , and $\epsilon > 0$, there exists δ_ϵ (depending possibly on the strategies) so that

$$\mathbb{E}_{\sigma^1, \sigma^2}^{\mu_0} (1 - \delta_\epsilon) \left[\sum_{m=0}^{\infty} \beta_\epsilon^m u^1(a_m^1, a_m^2) \right] + \epsilon \geq \liminf_{N \rightarrow \infty} \frac{1}{N} \mathbb{E}_{\sigma^1, \sigma^2}^{\mu_0} \left[\sum_{m=0}^{N-1} u^1(a_m^1, a_m^2) \right]$$

Now, let σ_n^1, σ_n^2 be a sequence of strategies which converge to the supremum for the average payoff. Let $\tilde{\sigma}_n^1, \tilde{\sigma}_n^2$ be one which comes within $\epsilon/2$ of the supremum so that

$$\begin{aligned} &\sup_{\sigma^1, \sigma^2} \limsup_{N \rightarrow \infty} \frac{1}{N} \mathbb{E}_{\sigma^1, \sigma^2} \left[\sum_{m=0}^{N-1} u^1(a_m^1, a_m^2) \right] \\ &\leq \limsup_{N \rightarrow \infty} \frac{1}{N} \mathbb{E}_{\tilde{\sigma}_n^1, \tilde{\sigma}_n^2} \left[\sum_{m=0}^{N-1} u^1(a_m^1, a_m^2) \right] + \epsilon/2 \end{aligned}$$

Let now δ_ϵ close to 1 be a discount factor whose optimal comes within $\epsilon/2$ of the limit when $\delta = 1$. For this parameter, under $\tilde{\sigma}_n^1, \tilde{\sigma}_n^2$ one obtains an upper bound on this payoff, which can be further upper bounded by optimizing over all possible strategies for this δ_ϵ value. This leads to a stationary strategy. Thus,

$$\begin{aligned} &\sup_{\sigma^1, \sigma^2} \limsup_{N \rightarrow \infty} \frac{1}{N} \mathbb{E}_{\sigma^1, \sigma^2} \left[\sum_{m=0}^{N-1} u^1(a_m^1, a_m^2) \right] - \epsilon/2 \\ &\leq \limsup_{N \rightarrow \infty} \frac{1}{N} \mathbb{E}_{\tilde{\sigma}_n^1, \tilde{\sigma}_n^2} \left[\sum_{m=0}^{N-1} u^1(a_m^1, a_m^2) \right] \\ &\leq \mathbb{E}_{\tilde{\sigma}_n^1, \tilde{\sigma}_n^2} (1 - \delta_\epsilon) \left[\sum_{m=0}^{\infty} \delta_\epsilon^m u^1(a_m^1, a_m^2) \right] + \epsilon/2 \\ &\leq \mathbb{E}_{\tilde{\sigma}^1, \tilde{\sigma}^2} (1 - \delta_\epsilon) \left[\sum_{m=0}^{\infty} \delta_\epsilon^m u^1(a_m^1, a_m^2) \right] + \epsilon/2 \\ &\leq \limsup_{N \rightarrow \infty} \frac{1}{N} \mathbb{E}_{\tilde{\sigma}^1, \tilde{\sigma}^2} \left[\sum_{m=0}^{N-1} u^1(a_m^1, a_m^2) \right] + \epsilon/2 + \epsilon', \end{aligned} \quad (16)$$

where ϵ' is a consequence of the following analysis. Under any stationary optimal strategy $\tilde{\sigma}^1, \tilde{\sigma}^2$ for a discounted problem,

$$\mathbb{E}_{\tilde{\sigma}^1, \tilde{\sigma}^2}(1 - \delta_\epsilon) \left[\sum_{m=0}^{\infty} \delta_\epsilon^m u^1(a_m^1, a_m^2) \right] - \limsup_{N \rightarrow \infty} \frac{1}{N} \mathbb{E}_{\tilde{\sigma}^1, \tilde{\sigma}^2} \left[\sum_{m=0}^{N-1} u^1(a_m^1, a_m^2) \right] \quad (17)$$

is uniformly bounded over all stationary policies under Assumption 4.1. Thus, one can select ϵ' and then ϵ arbitrarily small so that the result holds in the following fashion: First pick $\epsilon' > 0$, find a corresponding $\delta_{\epsilon'}$ with the understanding that for all $\delta_\epsilon \in [\delta_{\epsilon'}, 1)$, (16) holds. Now select $\delta_\epsilon \geq \delta_{\epsilon'}$ to satisfy the second inequality, such a δ_ϵ is guaranteed to exist since there are infinitely many such δ values up to 1 that satisfies this inequality. Here the uniformity of the convergence in (17) over all stationary policies is crucial.

In the above analysis, $\tilde{\sigma}^1, \tilde{\sigma}^2$ are stationary and with this stationary strategy,

$$\lim_{N \rightarrow \infty} \frac{1}{N} \mathbb{E}_{\mu_0}^{\mu^1, \mu^2} \left[\sum_{m=0}^{N-1} u^1(a_m^1, a_m^2) \right] \rightarrow \int \nu^*(d\mu, \gamma) G(\mu, \gamma)$$

by the convergence of the expected empirical occupation measures, where ν^* is the invariant probability measure which has full support on the Stackelberg strategies.

This leads to the following result which say that the infimum over all strategies is equal to the infimum over strategies which satisfy the structure given in Lemma 3.3, let us call such strategies μ_M :

$$\begin{aligned} & \inf_{\sigma^1} \liminf_{N \rightarrow \infty} \frac{1}{N} \mathbb{E}_{\sigma^1, \sigma^2}^{\mu_0} \sum_{m=0}^{N-1} u^1(a_m^1, a_m^2) \\ &= \inf_{\sigma^1 \in \mu_M} \liminf_{N \rightarrow \infty} \frac{1}{N} \mathbb{E}_{\sigma^1, \sigma^2}^{\mu_0 = \mu^*} \sum_{m=0}^{N-1} u^1(a_m^1, a_m^2) \end{aligned} \quad (18)$$

Finally, we establish the following:

$$\begin{aligned} & \inf_{\sigma^1} \liminf_{N \rightarrow \infty} \frac{1}{N} \mathbb{E}_{\sigma^1, \sigma^2}^{\mu_0} \sum_{m=0}^{N-1} u^1(a_m^1, a_m^2) \\ &= \inf_{\sigma^1 \in \mu_M} \liminf_{N \rightarrow \infty} \frac{1}{N} \mathbb{E}_{\sigma^1, \sigma^2}^{\mu_0} \sum_{m=0}^{N-1} u^1(a_m^1, a_m^2) \end{aligned} \quad (19)$$

This follows from the fact that,

$$\begin{aligned} & \inf_{\sigma^1} \liminf_{N \rightarrow \infty} \frac{1}{N} \mathbb{E}_{\sigma^1, \sigma^2}^{\mu_0} \sum_{m=0}^{N-1} u^1(a_m^1, a_m^2) \\ & \geq \inf_{\sigma^1 \in \mu_M} \liminf_{N \rightarrow \infty} \frac{1}{N} \mathbb{E}_{\sigma^1, \sigma^2}^{\mu_0 = \mu^*} \sum_{m=0}^{N-1} u^1(a_m^1, a_m^2) \end{aligned} \quad (20)$$

and that by the identifiability condition through using the Stackelberg strategies optimal for $\mu_0 = \mu^*$ to an arbitrary μ_0 , one obtains that

$$\begin{aligned} & \inf_{\sigma^1} \liminf_{N \rightarrow \infty} \frac{1}{N} \mathbb{E}_{\sigma^1, \sigma^2}^{\mu_0} \sum_{m=0}^{N-1} u^1(a_m^1, a_m^2) \\ & - \inf_{\sigma^1 \in \mu_M} \liminf_{N \rightarrow \infty} \frac{1}{N} \mathbb{E}_{\sigma^1, \sigma^2}^{\mu_0 = \mu^*} \sum_{m=0}^{N-1} u^1(a_m^1, a_m^2) \\ & = 0 \end{aligned} \tag{21}$$

On the other hand, once a strategy is given in μ_M , due to the identifiability assumption, any optimal strategy will need to be infinite repetition of a stage game Stackelberg action. \square

A.6 Proof of Lemma 5.1.

Suppose that $\max_x u^2(a^1, x) = u^2(a^1, x^*)$. Let $P_\sigma(a^1 | z_{[0,t]}^2) \geq 1 - \epsilon$. Let the maximum of

$$P_\sigma(a^1 | z_{[0,t]}^2) u^2(a^1, x) + \sum_{\bar{a}_j^1 \neq a^1} P_\sigma(\bar{a}_j^1 | z_{[0,t]}^2) u^2(\bar{a}_j^1, x)$$

be achieved by x^* so that

$$\begin{aligned} & P_\sigma(a^1 | z_{[0,t]}^2) u^2(a^1, x') + \sum_{\bar{a}_j^1 \neq a^1} P_\sigma(\bar{a}_j^1 | z_{[0,t]}^2) u^2(\bar{a}_j^1, x') \\ & \leq P_\sigma(a^1 | z_{[0,t]}^2) u^2(a^1, x^*) + \sum_{\bar{a}_j^1 \neq a^1} P_\sigma(\bar{a}_j^1 | z_{[0,t]}^2) u^2(\bar{a}_j^1, x^*) \end{aligned}$$

for any x' . For this to hold it suffices that

$$P_\sigma(a^1 | z_{[0,t]}^2) (u^2(a^1, x^*) - u^2(a^1, x')) \geq \max_{s,t} \epsilon u^2(s, t)$$

and since $P_\sigma(a^1 | z_{[0,t]}^2) \geq 1 - \epsilon$

$$(u^2(a^1, x^*) - u^2(a^1, x')) \geq \frac{\max_{s,t} \epsilon u^2(s, t)}{1 - \epsilon}$$

If $P_\sigma(a^1 | z_{[0,t]}^2) \geq 1 - \epsilon$ and for all $\bar{a}_j^1 \neq a^1$, $P_\sigma(\bar{a}_j^1 | z_{[0,t]}^2) \leq \epsilon/n$, (8) holds. \square

A.7 Proof of Theorem 5.1.

(8) is equivalent to, by Bayes' rule:

$$\frac{P_\sigma(z_{[0,t]}^2 | \hat{\omega} = m)}{P_\sigma(z_{[0,t]}^2 | \hat{\omega} = k)} \geq \frac{P_\sigma(\hat{\omega} = k) f(n)}{P_\sigma(\hat{\omega} = m)}$$

and

$$\sum_{j=0}^n \log \frac{(P_\sigma(z_j^2|\hat{\omega} = m))}{(P_\sigma(z_j^2|\hat{\omega} = k))} \geq \log \left(\frac{P_\sigma(\hat{\omega} = k)f(n)}{P_\sigma(\hat{\omega} = m)} \right)$$

Note now that (8) implies that $\tau_N^\omega \leq t$. Thus, we can now apply a measure concentration result through McDiarmid's inequality (see Raginsky and Sason [48]) to deduce that

$$\begin{aligned} & P_\sigma(\tau_N \geq t) \\ & \leq P \left(\sum_{j=0}^t \log \left(\frac{P_\sigma(z_j^2|\hat{\omega} = m)}{P_\sigma(z_j^2|\hat{\omega} = k)} \right) \leq \log \left(\frac{P_\sigma(\hat{\omega} = k)f(n)}{P_\sigma(\hat{\omega} = m)} \right) \right) \\ & \leq P \left(\frac{1}{t+1} \sum_{j=0}^t \log \left(\frac{P_\sigma(z_j^2|\hat{\omega} = m)}{P_\sigma(z_j^2|\hat{\omega} = k)} \right) - \mathbb{E} \left[\log \frac{P_\sigma(z_j^2|\hat{\omega} = m)}{P_\sigma(z_j^2|\hat{\omega} = k)} \right] \right. \\ & \quad \left. \leq \frac{1}{t+1} \log \left(\frac{P_\sigma(\hat{\omega} = k)f(n)}{P_\sigma(\hat{\omega} = m)} \right) - \mathbb{E} \left[\log \frac{P_\sigma(z_j^2|\hat{\omega} = m)}{P_\sigma(z_j^2|\hat{\omega} = k)} \right] \right) \\ & \leq P \left(\left| \frac{1}{t+1} \sum_{j=0}^t \log \left(\frac{P_\sigma(z_j^2|\hat{\omega} = m)}{P_\sigma(z_j^2|\hat{\omega} = k)} \right) - \mathbb{E} \left[\log \frac{P_\sigma(z_j^2|\hat{\omega} = m)}{P_\sigma(z_j^2|\hat{\omega} = k)} \right] \right| \right. \\ & \quad \left. \geq \left| \mathbb{E} \left[\log \frac{P_\sigma(z_j^2|\hat{\omega} = m)}{P_\sigma(z_j^2|\hat{\omega} = k)} \right] - \frac{1}{t+1} \log \left(\frac{P_\sigma(\hat{\omega} = k)f(n)}{P_\sigma(\hat{\omega} = m)} \right) \right| \right) \\ & \leq 2e^{-n \left(\mathbb{E} \left[\log \frac{P_\sigma(z_j^2|\hat{\omega} = m)}{P_\sigma(z_j^2|\hat{\omega} = k)} \right] - \frac{1}{t+1} \log \left(\frac{P_\sigma(\hat{\omega} = k)f(n)}{P_\sigma(\hat{\omega} = m)} \right) \right)^2 / (b-a)} \end{aligned} \tag{22}$$

where $a \leq \mathbb{Z}^j \leq b$ with $\mathbb{Z}^j = \frac{P_\sigma(z_j^2|\hat{\omega} = m)}{P_\sigma(z_j^2|\hat{\omega} = k)}$. This implies, since $\log(n)/n \rightarrow 0$ and $f(n) = n(1-\epsilon)/\epsilon$ and by Lemma 5.1, that the probability of $\tau_N \geq t$ is upper bounded asymptotically by a geometric random variable, that is, there exists $R < \infty$ and $\rho \in (0, 1)$ so that for all $k \in \mathbb{N}$, $P_\sigma(\tau_m \geq t) \leq R\rho^t$. \square

A.8 Proof of Lemma 6.1.

From (11), we observe the following. Let f be a continuous function on $\Delta(\Omega)$. Then $\int f(\mu_{t+1}|\mu_t, \gamma_t^1)$ is continuous in (μ_t, γ_t^1) if

$$\sum_{z_t^2} f(H(\mu_t, z_t^2, \gamma_t^1)) P_\sigma(z_t^2|\gamma_t^1)$$

is continuous in μ_t, γ_t^1 where $\mu_{t+1} = H(\mu_t, z_t^2, \gamma_t^1)$ defined by (11) with the variables

$$1_{\{\gamma_t^1(\omega, z_{[0,t-1]}^2) = a_t^1\}} = P_\sigma(a_t^1|\omega, z_{[0,t-1]}^2), \quad \mu_t(\omega) = P_\sigma(\omega|z_{[0,t-1]}^2)$$

Instead of considering continuous functions on $\Delta(\Omega)$, we can also consider continuity of $\mu_{t+1}(\omega)$ for every ω since pointwise convergence implies convergence in total variation and which in turn implies convergence under weak convergence by Scheffé's Theorem. Now, for every fixed $z_t^2 = z$, $\mu_{t+1}(\omega)$ is continuous in μ_t for every ω , and hence $H(\mu_t, z_t^2, \gamma_t^1)$ is

continuous in total variation since pointwise convergence implies convergence in total variation. Furthermore, $P_\sigma(z_t^2|\gamma_t^1, \mu_t)$ is continuous in μ_t for a given γ_t^1 ; thus, weak continuity follows. \square

A.9 Brief Review of Markov Decision Processes

We provide below a brief review of Markov Decision Processes. Note that the notation we employ is the standard notation in the optimal stochastic control theory.¹⁵

Let Z be a Borel space (i.e., a Borel subset of a complete and separable metric space) and let $\mathcal{P}(Z)$ denote the set of all probability measures on Z .

Definition A.1 (Markov Control Model [31]). *A discrete time Markov control model (Markov decision process) is a system characterized by the 4-tuple*

$$(Z, U, K, c),$$

where

1. Z is the state space, the set of all possible states of the system;
2. U (a Borel space) is the control space (or action space), the set of all controls (actions) $a \in U$ that can act on the system;
3. $K = K(\cdot|z, a)$ is the transition probability of the system, a stochastic kernel on Z given $Z \times U$, i.e., $K(\cdot|z, a)$ is a probability measure on Z for all state-action pairs (z, a) , and $K(B|\cdot, \cdot)$ is a measurable function from $Z \times U$ to $[0, 1]$ for each Borel set $B \subset Z$;
4. $c : Z \times U \rightarrow [0, \infty)$ is the cost per time stage function of the system, a function $c(x, a)$ of the state and the control.

Define the *history* spaces H_t at time $t \geq 0$ of the Markov control model by $H_0 := Z$ and $H_t := (Z \times U)^t \times Z$. Thus a specific history $h_t \in H_t$ has the form $h_t = (z_0, u_0, \dots, z_{t-1}, u_{t-1}, z_t)$.

Definition A.2 (Admissible Control Policy [31]). *An admissible control policy $\Pi = \{\alpha_t\}_{t \geq 0}$, also called a randomized control policy (more simply a control policy or a policy) is a sequence of stochastic kernels on the action space U given the history H_t . The set of all randomized control policies is denoted by Π_A . A deterministic policy Π is a sequence of functions $\{\alpha_t\}_{t \geq 0}$, $\alpha_t : H_t \rightarrow U$, that determine the control used at each time stage deterministically, i.e., $a_t = \alpha_t(h_t)$. The set of all deterministic policies is denoted Π_D . Note that $\Pi_D \subset \Pi_A$. A Markov policy is a policy Π such that for each time stage the choice of control only depends on the current state z_t , i.e., $\Pi = \{\alpha_t\}_{t \geq 0}$ with $\alpha_t : Z \rightarrow \mathcal{P}(U)$. The set of all Markov policies is denoted by Π_M . The set of deterministic Markov policies is denoted by Π_{MD} . A stationary policy is a Markov policy $\Pi = \{\alpha_t\}_{t \geq 0}$ such that $\alpha_t = \alpha$ for all $t \geq 0$ for some $\alpha : Z \rightarrow \mathcal{P}(U)$. The set of all stationary policies is denoted by Π_S and the set of deterministic stationary policies is denoted by Π_{SD} .*

¹⁵In the standard stochastic control theory, the decision maker is to minimize a cost function rather than maximizing a payoff function. Needless to say, payoff function maximizing analogs of the results presented here can be obtained trivially.

According to the Ionescu Tulcea theorem (see [31]), the transition kernel K , an initial probability distribution μ_0 on \mathbf{Z} , and a policy Π define a unique probability measure $P_{\mu_0}^\Pi$ on $\mathbf{H}_\infty = (\mathbf{X} \times \mathbf{U})^\infty$, the distribution of the state-action process $\{(Z_t, A_t)\}_{t \geq 0}$. The resulting state process $\{Z_t\}_{t \geq 0}$ is called a *controlled Markov process*. The expectation with respect to $P_{\mu_0}^\Pi$ is denoted by $E_{\mu_0}^\Pi$. If $\mu_0 = \delta_z$, the point mass at $z \in \mathbf{Z}$, we write P_z^Π and E_z^Π instead of $P_{\delta_z}^\Pi$ and $E_{\delta_z}^\Pi$.

In an *optimal control problem*, a performance objective J of the system is given and the goal is to find the controls that minimize (or maximize) that objective. Some common optimal control problems for Markov control models are the following:

1. *Finite Horizon Average Cost Problem*: Here the goal is to find policies that minimize the average cost

$$J_{\mu_0}(\Pi, T) := E_{\mu_0}^\Pi \left[\frac{1}{T} \sum_{t=0}^{T-1} c(Z_t, U_t) \right], \quad (23)$$

for some $T \geq 1$.

2. *Infinite Horizon Discounted Cost Problem*: Here the goal is to find policies that minimize

$$J_{\mu_0}^\beta(\Pi) := \lim_{T \rightarrow \infty} E_{\mu_0}^\Pi \left[\sum_{t=0}^{T-1} \beta^t c(Z_t, U_t) \right], \quad (24)$$

for some $\beta \in (0, 1)$.

3. *Infinite Horizon Average Cost Problem*: In the more challenging infinite horizon control problem the goal is to find policies that minimize the average cost

$$J_{\mu_0}(\Pi) := \limsup_{T \rightarrow \infty} E_{\mu_0}^\Pi \left[\frac{1}{T} \sum_{t=0}^{T-1} c(Z_t, U_t) \right]. \quad (25)$$

The Markov control model together with the performance objective is called a *Markov decision process*.

A common method to solving finite horizon Markov control problems is by *dynamic programming*, which involves working backwards from the final time stage to solve for the optimal sequence of controls to use. The optimality of this algorithm is guaranteed by Bellman's principle of optimality.

Theorem A.1 (Bellman's Principle of Optimality [31, Chapter 3.2]). *Given a finite time horizon $T \geq 1$, define a sequence of functions J_T, \dots, J_0 on \mathbf{Z} recursively such that*

$$J_T(z_T) \equiv 0,$$

and for $0 \leq t < T$ and $z \in \mathbf{Z}$,

$$J_t(z) := \min_{a \in \mathbf{U}} \left[c(z, a) + \int_{\mathbf{Z}} J_{t+1}(z') K(dz' | z, a) \right]. \quad (26)$$

If the J_t are measurable and there exist measurable $f_t : \mathbf{Z} \rightarrow \mathbf{U}$ such that $a = f_t(z)$ achieves the above minimum for all $t = 0, \dots, T-1$, then the deterministic Markov policy $\Pi := (f_0, \dots, f_{T-1})$ is optimal with cost $J_{z_0}(\Pi, T) = J_0(z_0)$.

General sufficient conditions exist under which the two assumptions of the above theorem hold [31, Chapter 3.3].

A stochastic kernel K on Z given $Z \times U$ is called weakly continuous if the function $(a, z) \mapsto \int_Z v(z')K(dz'|z, a)$ is continuous whenever v is a *bounded and continuous* real function on $Z \times U$. It is called strongly continuous if the $(a, z) \mapsto \int_Z v(z')K(dz'|z, a)$ is continuous whenever v is a *measurable and bounded* real function on $Z \times U$. The next theorems follow from [31, Chapter 3.3] and [32, Chapter 8.5].

For the infinite horizon discounted cost Markov control problem, one can use an iteration algorithm to obtain an optimal policy. This approach is commonly called the *successive approximations* or value iteration method [31, Chapter 4.2].

Theorem A.2. *Suppose the following conditions hold:*

- (i) *The one stage cost c is continuous, nonnegative, and bounded;*
- (ii) *U is compact;*
- (iii) *the transition kernel K is weakly continuous.*

Then for any $\beta \in (0, 1)$, the pointwise limit $J(z)$ as $t \rightarrow \infty$, of the sequence defined by

$$J_t(z) = \min_{u \in U} \left[c(z, u) + \beta \int_Z J_{t-1}(z')K(z'|z, u) \right], \quad z \in Z,$$

with $J_0(z) \equiv 0$, yields the optimum cost in the infinite horizon discounted cost problem (i.e., $\inf_{\Pi \in \Pi_A} J_z^\beta = J(z)$). Furthermore, there exists a measurable function $f : Z \rightarrow U$ such that

$$J(z) = c(z, f(z)) + \beta \int_Z J_{t-1}(z')K(z'|z, f(z))$$

and the policy $\Pi = \{f\}$ is an optimal stationary Markov policy. Furthermore, $J(z)$ is continuous.

Theorem A.3. *Suppose the following conditions hold:*

- (i) *The one stage cost c is nonnegative, and bounded and continuous in u for every z ;*
- (ii) *U is compact;*
- (iii) *the transition kernel K is strongly continuous in u for every x .*

Then for any $\beta \in (0, 1)$, the pointwise limit $J(z)$ as $t \rightarrow \infty$, of the sequence defined by

$$J_t(z) = \min_{u \in U} \left[c(z, u) + \beta \int_Z J_{t-1}(z')K(z'|z, u) \right], \quad z \in Z,$$

with $J_0(z) \equiv 0$, yields the optimum cost in the infinite horizon discounted cost problem (i.e., $\inf_{\Pi \in \Pi_A} J_z^\beta = J(z)$). Furthermore, there exists a measurable function $f : Z \rightarrow U$ such that

$$J(z) = c(x, f(z)) + \beta \int_Z J_{t-1}(z')K(z'|z, f(z))$$

and the policy $\Pi = \{f\}$ is an optimal stationary Markov policy.

References

- [1] D. Abreu, D. Pearce, and E. Stacchetti. Toward a theory of discounted repeated games with imperfect monitoring. *Econometrica*, 58(5):1041–1063, 1990.
- [2] A. Atakan and M. Ekmekci. Reputation in long-run relationships. *Review of Economic Studies*, 79(2):451–480, 2012.
- [3] A. Atakan and M. Ekmekci. A two-sided reputation result with long-run players. *Journal of Economic Theory*, 148(1):376–392, 2013.
- [4] A. Atakan and M. Ekmekci. Reputation in the long-run with imperfect monitoring. forthcoming *Journal of Economic Theory*.
- [5] R.J. Barro. Reputation in a model of monetary policy with incomplete information. *Journal of Monetary Economics*, 17:3–20, 1986.
- [6] J. Bergin. A characterization of sequential equilibrium strategies in infinitely repeated incomplete information games. *Journal of Economic Theory*, 47(1):51–65, 1989.
- [7] V. S. Borkar. *Probability theory: an advanced course*. Springer, 2012.
- [8] V. S. Borkar. White-noise representations in stochastic realization theory. *SIAM J. on Control and Optimization*, 31:1093–1102, 1993.
- [9] V. S. Borkar, S. K. Mitter, and S. Tatikonda. Optimal sequential vector quantization of Markov sources. *SIAM J. Control and Optimization*, 40:135–148, 2001.
- [10] M. Celentani and W. Pesendorfer. Reputation in dynamic games. *Journal of Economic Theory*, 70:109–132, 1996.
- [11] V.V. Chari and P.J. Kehoe. Sustainable plans and debt. *Journal of Economic Theory*, 61:230–261, 1993.
- [12] H.L. Cole, J. Dow, and W.B. English. Default, settlement, and signalling: Lending resumption in a reputational model of sovereign debt. *International Economic Review*, 36:365–385, 1995.
- [13] M.W. Cripps and E. Faingold. The value of a reputation under imperfect monitoring. mimeo, 2015.
- [14] M.W. Cripps, G.J. Mailath, and L. Samuelson. Imperfect monitoring and impermanent reputations. *Econometrica*, 72(2):407–432, 2004.
- [15] A. Cukierman and A. Meltzer. A theory of ambiguity, credibility and inflation under discretion and asymmetric information. *Econometrica*, 54:1099–1128, 1986.
- [16] N.A. Dalkiran. Order of limits in reputations. *Theory and Decision*, 81(3): 393– 411, 2016.
- [17] P. Diamond. Reputation acquisition in debt markets. *Journal of Political Economy*, 97:828–868, 1989.

- [18] L. Epstein, J. Noor, and A. Sandroni. Non-Bayesian Learning. *The B.E. Journal of Theoretical Economics*, 10(1):Article 3, 2010.
- [19] M. Ekmekci. Sustainable reputations with rating systems. *Journal of Economic Theory*, 146(2):479–503, 2011.
- [20] M. Ekmekci, O. Gossner, and A. Wilson. Impermanent types and permanent reputation. *Journal of Economic Theory*, 147(1):162–178, 2012.
- [21] J. Ely and J. Valimaki. Bad reputation. *Quarterly Journal of Economics*, 118:785–814, 2003.
- [22] E. Faingold. Reputation and the flow of information in repeated games. mimeo, 2014.
- [23] E. Faingold and Y. Sannikov. Reputation in continuous-time games. *Econometrica*, 79(3):773–876, 2011.
- [24] D. Fudenberg, D. M. Kreps, and E. Maskin. Repeated games with long-run and short-run players. *Review of Economic Studies*, 57:555–573, 1990.
- [25] D. Fudenberg and D.K. Levine. Reputation and equilibrium selection in games with a patient player. *Econometrica*, 57(4):759–778, 1989.
- [26] D. Fudenberg and D.K. Levine. Maintaining a reputation when strategies are imperfectly observed. *Review of Economic Studies*, 59(3):561–579, 1992.
- [27] D. Fudenberg and D.K. Levine. Efficiency and observability with long-run and short-run players. *Journal of Economic Theory*, 62(1):103–135, 1994.
- [28] D. Fudenberg, D.K. Levine, and E. Maskin. The folk theorem with imperfect public information. *Econometrica*, 62(5):997–1039, 1994.
- [29] D. Fudenberg and E. Maskin. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54(3):533–554, 1986.
- [30] O. Gossner. Simple bounds on the value of a reputation. *Econometrica*, 79:1627–1641, 2011.
- [31] O. Hernandez-Lerma and J. Lasserre. *Discrete-time Markov control processes*. Springer, 1996.
- [32] O. Hernández-Lerma and J.B. Lasserre. *Further Topics on Discrete-Time Markov Control Processes*. Springer, 1999.
- [33] J. Hörner. Reputation and competition. *American Economic Review*, 92(3):644–663, 2002.
- [34] J. Hörner and S. Lovo. Belief-free equilibria in games with incomplete information. *Econometrica*, 77(2):453–487, 2009.
- [35] B. Jullien and I.-U. Park. New, like new, or very good? reputation and credibility. *Review of Economic Studies*, 81(4):1543–1574, 2014.

- [36] E. Kalai and E. Lehrer. Rational learning leads to nash equilibrium. *Econometrica*, 61(5):1019–1045, 1993.
- [37] E. Kalai and E. Lehrer. Weak and strong merging of opinions. *Journal of Mathematical Economics*, 23(1):73–86, 1994.
- [38] B. Klein and K. B. Leffler. The role of market forces in assuring contractual performance. *Journal of Political Economy*, 89:615–641, 1981.
- [39] D. M. Kreps, P. Milgrom, D. Roberts, and R. Wilson. Rational cooperation in the finitely repeated prisoners’ dilemma. *Journal of Economic Theory*, 27(2):245–252, 1982.
- [40] D. M. Kreps and R. Wilson. Reputation and imperfect information. *Journal of Economic Theory*, 27(2):253–279, 1982.
- [41] T. Linder and S. Yüksel. On optimal zero-delay quantization of vector Markov sources. *IEEE Transactions on Information Theory*, 60:2975–5991, October 2014.
- [42] Q. Liu. Information acquisition and reputation dynamics. *Review of Economic Studies*, 78(4):1400–1425, 2011.
- [43] Q. Liu and A. Skrzypacz. Limited records and reputation bubbles. *Journal of Economic Theory*, 151:2–29, 2014.
- [44] A. Mahajan and D. Teneketzis. On the design of globally optimal communication strategies for real-time noisy communication with noisy feedback. *IEEE Journal on Selected Areas in Communications*, 26:580–595, May 2008.
- [45] P. Milgrom and D. Roberts. Predation, reputation and entry deterrence. *Journal of Economic Theory*, 27(2):280–312, 1982.
- [46] A. Özdoğan. Disappearance of reputations in two-sided incomplete-information games. *Games and Economic Behavior*, 88:211–220, 2014.
- [47] C. Phelan. Public trust and government betrayal. *Journal of Economic Theory*, 130:27–43, 2006.
- [48] M. Raginsky and I. Sason. Concentration of measure inequalities in information theory, communications and coding. *arXiv preprint arXiv:1212.4663*, 2012.
- [49] C.A. Sims. Implications of rational inattention. *Journal of Monetary Economics*, 50(3):665–690, 2003.
- [50] C.A. Sims. Rational inattention: Beyond the linear-quadratic case. *American Economic Review*, 96(2):158–163, 2006.
- [51] S. Sorin. Merging, reputation, and repeated games with incomplete information. *Games and Economic Behavior*, 29:274–308, 1999.
- [52] S. Tadelis. What’s in a name? reputation as a tradeable asset. *American Economic Review*, 89(3):548–563, 1999.

- [53] D. Teneketzis. On the structure of optimal real-time encoders and decoders in noisy communication. *IEEE Transactions on Information Theory*, 52:4017–4035, September 2006.
- [54] J. C. Walrand and P. Varaiya. Optimal causal coding-decoding problems. *IEEE Transactions on Information Theory*, 19:814–820, November 1983.
- [55] H. S. Witsenhausen. On the structure of real-time source coders. *Bell Syst. Tech. J.*, 58:1437–1451, July/August 1979.
- [56] S. Yüksel. On optimal causal coding of partially observed Markov sources in single and multi-terminal settings. *IEEE Transactions on Information Theory*, 59:424–437, January 2013.
- [57] S. Yüksel and T. Başar. *Stochastic Networked Control Systems: Stabilization and Optimization under Information Constraints*. Birkhäuser, Boston, MA, 2013.